

1. [Preface to the Chemistry of Electronic Materials](#)
2. Background to Electronic Materials
 1. [Introduction to Semiconductors](#)
 2. [Doped Semiconductors](#)
 3. [Diffusion](#)
 4. [Crystal Structure](#)
 5. [Structures of Element and Compound Semiconductors](#)
3. Device Fundamentals
 1. [Introduction to Bipolar Transistors](#)
 2. [Basic MOS Structure](#)
 3. [Introduction to the MOS Transistor and MOSFETs](#)
 4. [Light Emitting Diode](#)
 5. [Polymer Light Emitting Diodes](#)
 6. [Laser](#)
 7. [Solar Cells](#)
4. Bulk Materials
 1. [Properties of Gallium Arsenide](#)
 2. [Synthesis and Purification of Bulk Semiconductors](#)
 3. [Growth of Gallium Arsenide Crystals](#)
 4. [Ceramic Processing of Alumina](#)
 5. [Piezoelectric Materials Synthesis](#)
5. Wafer Formation and Processing
 1. [Formation of Silicon and Gallium Arsenide Wafers](#)
 2. [Doping](#)
 3. [Applications for Silica Thin Films](#)
 4. [Oxidation of Silicon](#)
 5. [Photolithography](#)
 6. [Optical Issues in Photolithography](#)
 7. [Composition and Photochemical Mechanisms of Photoresists](#)
 8. [Integrated Circuit Well and Gate Creation](#)

9. [Applying Metallization by Sputtering](#)
6. Thin Film Growth
 1. [Molecular Beam Epitaxy](#)
 2. [Atomic Layer Deposition](#)
 3. [Chemical Vapor Deposition](#)
 4. [Liquid Phase Deposition](#)
7. Chemical Vapor Deposition
 1. [Selecting a Molecular Precursor for Chemical Vapor Deposition](#)
 2. [Determination of Sublimation Enthalpy and Vapor Pressure for Inorganic and Metal-Organic Compounds by Thermogravimetric Analysis](#)
 3. 13-15 (III-V) Semiconductor Chemical Vapor Deposition
 1. [Phosphine and Arsine](#)
 2. [Mechanism of the Metal Organic Chemical Vapor Deposition of Gallium Arsenide](#)
 4. Oxide Chemical Vapor Deposition
 1. [Chemical Vapor Deposition of Silica Thin Films](#)
 2. [Chemical Vapor Deposition of Alumina](#)
 5. Nitride Chemical Vapor Deposition
 1. [Introduction to Nitride Chemical Vapor Deposition](#)
 2. [Chemical Vapor Deposition of Silicon Nitride and Oxynitride](#)
 3. [Chemical Vapor Deposition of Aluminum Nitride](#)
 6. [Metal Organic Chemical Vapor Deposition of Calcium Fluoride](#)
 7. [Precursors for Chemical Vapor Deposition of Copper](#)
8. Materials Characterization
 1. [Rutherford Backscattering of Thin Films](#)
 2. [The Application of VSI \(Vertical Scanning Interferometry\) to the Study of Crystal Surface Processes](#)
 3. [Atomic Force Microscopy](#)

9. Nanotechnology

1. [Introduction to Nanoparticle Synthesis](#)

2. Semiconductor Nanomaterials

1. [Synthesis of Semiconductor Nanoparticles](#)

2. [Optical Properties of Group 12-16 \(II-VI\) Semiconductor Nanoparticles](#)

3. [Characterization of Group 12-16 \(II-VI\) Semiconductor Nanoparticles by UV-visible Spectroscopy](#)

4. [Optical Characterization of Group 12-16 \(II-VI\) Semiconductor Nanoparticles by Fluorescence Spectroscopy](#)

3. [Carbon Nanomaterials](#)

4. [Graphene](#)

5. [Rolling Molecules on Surfaces Under STM Imaging](#)

10. Economic and Environmental Issues

1. [The Environmental Impact of the Manufacturing of Semiconductors](#)

Preface to the Chemistry of Electronic Materials

The intention of this text is not to provide a comprehensive reference to all aspects of semiconductor device fabrication or a review of research results that, irrespective of their promise, have not been adopted into mainstream production. Instead it is aimed to provide a useful reference for those interested in the chemical aspects of the electronics industry.

Given the nature of Connexions, this course is fluid in structure and content. In addition, it contains modules by other authors where appropriate. The content will be updated and expanded with time. If any authors have suitable content, please contact me and I will be glad to assist in transforming the content to a suitable module structure.

Andrew R. Barron

Rice University, Houston, TX 77005. E-mail: arb@rice.edu

Introduction to Semiconductors

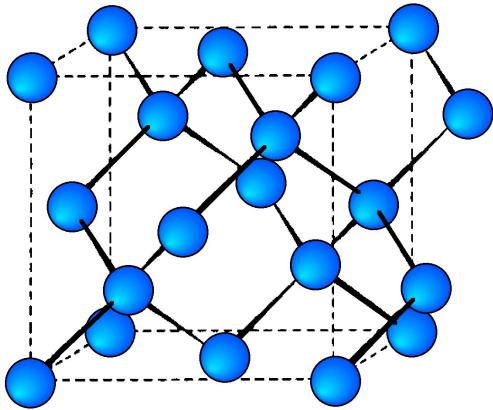
Introduction to semiconductors, mainly looking at the behavior of electrons in a solid from a quantum mechanical point of view.

Note: This module is adapted from the Connexions module entitled *Introduction to Semiconductors* by Bill Wilson.

If we only had to worry about simple conductors, life would not be very complicated, but on the other hand we wouldn't be able to make computers, CD players, cell phones, i-Pods and a lot of other things which we have found to be useful. We will now move on, and talk about another class of conductors called semiconductors.

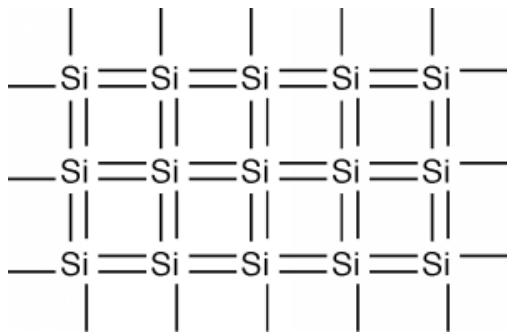
In order to understand semiconductors and in fact to get a more accurate picture of how metals, or normal conductors actually work, we really have to resort to quantum mechanics. Electrons in a solid are very tiny objects, and it turns out that when things get small enough, they no longer exactly following the classical "Newtonian" laws of physics that we are all familiar with from everyday experience. It is not the purpose of this course to teach quantum mechanics, so what we are going to do instead is describe the results which come from looking at the behavior of electrons in a solid from a quantum mechanical point of view.

Solids (at least the ones we will be talking about, and especially semiconductors) are crystalline materials, which means that they have their atoms arranged in a ordered fashion. We can take silicon (the most important semiconductor) as an example. Silicon is a group 14(IV) element, which means it has four electrons in its outer or valence shell. Silicon crystallizes in a structure called the diamond crystal lattice, shown in [\[link\]](#). Each silicon atom has four covalent bonds, arranged in a tetrahedral formation about the atom center.



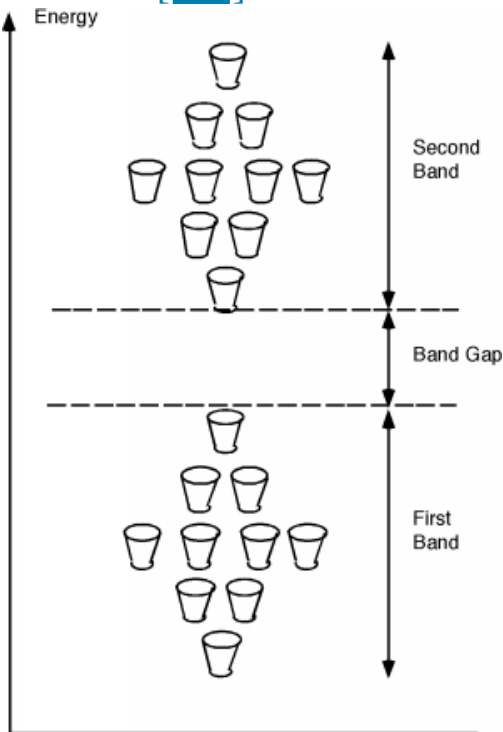
Crystal structure of silicon.

In two dimensions, we can schematically represent a piece of single-crystal silicon as shown in [\[link\]](#). Each silicon atom shares its four valence electrons with valence electrons from four nearest neighbors, filling the shell to 8 electrons, and forming a stable, periodic structure. Once the atoms have been arranged like this, the outer valence electrons are no longer strongly bound to the host atom. The outer shells of all of the atoms blend together and form what is called a band. The electrons are now free to move about within this band, and this can lead to electrical conductivity as we discussed earlier.



A 2-D representation of a silicon crystal.

This is not the complete story however, for it turns out that due to quantum mechanical effects, there is not just one band which holds electrons, but several of them. What will follow is a very qualitative picture of how the electrons are distributed when they are in a periodic solid, and there are necessarily some details which we will be forced to gloss over. On the other hand this will give you a pretty good picture of what is going on, and may enable you to have some understanding of how a semiconductor really works. Electrons are not only distributed throughout the solid crystal spatially, but they also have a distribution in energy as well. The potential energy function within the solid is periodic in nature. This potential function comes from the positively charged atomic nuclei which are arranged in the crystal in a regular array. A detailed analysis of how electron wave functions, the mathematical abstraction which one must use to describe how small quantum mechanical objects behave when they are in a periodic potential, gives rise to an energy distribution somewhat like that shown in [\[link\]](#).

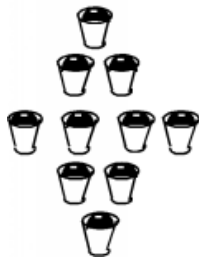


Schematic of the first two bands in a periodic solid showing energy levels and bands.

Firstly, unlike the case for free electrons, in a periodic solid, electrons are not free to take on any energy value they wish. They are forced into specific energy levels called allowed states, which are represented by the cups in [\[link\]](#). The allowed states are not distributed uniformly in energy either. They are grouped into specific configurations called energy bands. There are no allowed levels at zero energy and for some distance above that. Moving up from zero energy, we then encounter the first energy band. At the bottom of the band there are very few allowed states, but as we move up in energy, the number of allowed states first increases, and then falls off again. We then come to a region with no allowed states, called an energy band gap. Above the band gap, another band of allowed states exists. This goes on and on, with any given material having many such bands and band gaps. This situation is shown schematically in [\[link\]](#), where the small cups represent allowed energy levels, and the vertical axis represents electron energy.

It turns out that each band has exactly $2N$ allowed states in it, where N is the total number of atoms in the particular crystal sample we are talking about. (Since there are 10 cups in each band in the figure, it must represent a crystal with just 5 atoms in it. Not a very big crystal at all!) Into these bands we must now distribute all of the valence electrons associated with the atoms, with the restriction that we can only put one electron into each allowed state. This is the result of something called the Pauli exclusion principle. Since in the case of silicon there are 4 valence electrons per atom, we would just fill up the first two bands, and the next would be empty. If we make the logical assumption that the electrons will fill in the levels with the lowest energy first, and only go into higher lying levels if the ones below are already filled. This situation is shown in [\[link\]](#), in which we have represented electrons as small black balls with a "-" sign on them. Indeed, the first two bands are completely full, and the next is empty. What will happen if we apply an electric field to the sample of silicon? Remember the diagram we have at hand right now is an energy based one, we are showing how the electrons are distributed in energy, not how they are arranged spatially. On this diagram we can not show how they will move about, but only how they will change their energy as a result of the applied field. The

electric field will exert a force on the electrons and attempt to accelerate them. If the electrons are accelerated, then they must increase their kinetic energy. Unfortunately, there are no empty allowed states in either of the filled bands. An electron would have to jump all the way up into the next (empty) band in order to take on more energy. In silicon, the gap between the top of the highest most occupied band and the lowest unoccupied band is 1.1 eV. (One eV is the potential energy gained by an electron moving across an electrical potential of one volt.) The mean free path or distance over which an electron would normally move before it suffers a collision is only a few hundred angstroms (*ca.* 300×10^{-8} cm) and so you would need a very large electric field (several hundred thousand V/cm) in order for the electron to pick up enough energy to "jump the gap". This makes it appear that silicon would be a very bad conductor of electricity, and in fact, very pure silicon is very poor electrical conductor.



Silicon
, with
first
two

bands
full
and the
next
empty.

A metal is an element with an odd number of valence electrons so that a metal ends up with an upper band which is just half full of electrons. This is illustrated in [\[link\]](#). Here we see that one band is full, and the next is just half full. This would be the situation for the Group 13(III) element aluminum for instance. If we apply an electric field to these carriers, those near the top of the distribution can indeed move into higher energy levels by acquiring some kinetic energy of motion, and easily move from one place to the next. In reality, the whole situation is a bit more complex than we have shown here, but this is not too far from how it actually works.



Electron
distributio

n for a
metal or
good
conductor.

So, back to our silicon sample. If there are no places for electrons to "move" into, then how does silicon work as a "semiconductor"? Well, in the first place, it turns out that not all of the electrons are in the bottom two bands. In silicon, unlike say quartz or diamond, the band gap between the top-most full band, the next empty one is not so large. As we mentioned above it is only about 1.1 eV. So long as the silicon is not at absolute zero temperature, some electrons near the top of the full band can acquire enough thermal energy that they can "hop" the gap, and end up in the upper band, called the conduction band. This situation is shown in [\[link\]](#).



Thermal
excitation
of
electrons
across the
band gap.

In silicon at room temperature, roughly 10^{10} electrons per cubic centimeter are thermally excited across the band-gap at any one time. It should be noted that the excitation process is a continuous one. Electrons are being excited across the band, but then they fall back down into empty spots in the lower band. On average however, the 10^{10} in each cm^3 of silicon is what you will find at any given instant. Now 10 billion electrons per cubic centimeter seems like a lot of electrons, but lets do a simple calculation. The mobility of electrons in silicon is about $1000 \text{ cm}^2/\text{V.s}$. Remember, mobility times electric field yields the average velocity of the carriers. Electric field has units of V/cm , so with these units we get velocity in cm/s as we should. The charge on an electron is 1.6×10^{-19} coulombs. Thus from [\[link\]](#):

Equation:

$$\begin{aligned}\sigma &= nq\mu \\ &= 10^{10} (1.6 \times 10^{-19}) 1000 \\ &= 1.6 \times 10^{-6} \text{ mhos/cm}\end{aligned}$$

If we have a sample of silicon 1 cm long by (1 mm x 1mm) square, it would have a resistance, [\[link\]](#), which does not make it much of a "conductor". In fact, if this were all there was to the silicon story, we could pack up and move on, because at any reasonable temperature, silicon would conduct electricity very poorly.

Equation:

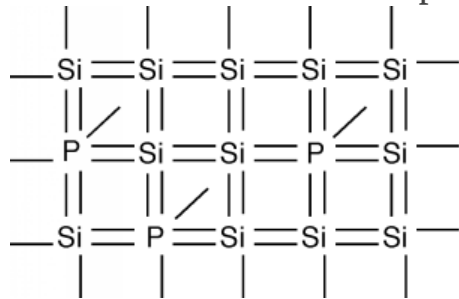
$$\begin{aligned}R &= L/\sigma A \\ &= 1/(1.6 \times 10^{-6})(0.1)^2 \\ &= 1.6 \times 10^{-6} \text{ M}\Omega\end{aligned}$$

Doped Semiconductors

From the silicon's crystal structure to discuss how to make doped semiconductors and the mechanics.

Note: This module is adapted from the Connexions module entitled *Doped Semiconductors* by Bill Wilson.

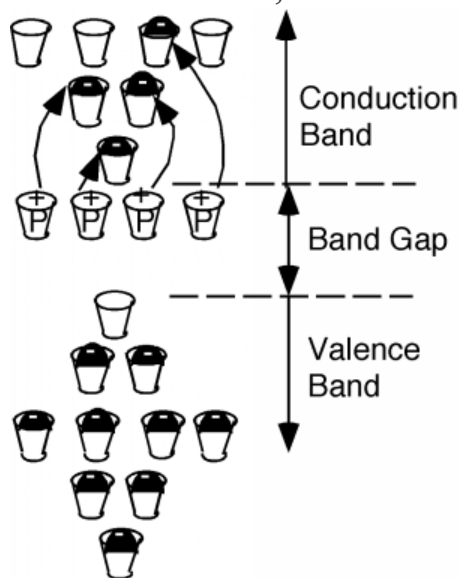
To see how we can make silicon a useful electronic material, we will have to go back to its crystal structure ([\[link\]](#)). Suppose somehow we could substitute a few atoms of phosphorus for some of the silicon atoms.



A two dimensional
representation of a
silicon crystal lattice
"doped" with
phosphorus.

If you sneak a look at the periodic table, you will see that phosphorus is a group V element, as compared with silicon which is a group 14(IV) element. What this means is the phosphorus atom has five outer or valence electrons, instead of the four which silicon has. In a lattice composed mainly of silicon, the extra electron associated with the phosphorus atom has no "mating" electron with which it can complete a shell, and so is left loosely dangling to the phosphorus atom, with relatively low binding energy. In fact, with the addition of just a little thermal energy (from the

natural or latent heat of the crystal lattice) this electron can break free and be left to wander around the silicon atom freely. In our "energy band" picture, we have something like what we see in [\[link\]](#). The phosphorus atoms are represented by the added cups with P's on them. They are new allowed energy levels which are formed within the "band gap" near the bottom of the first empty band. They are located close enough to the empty (or "conduction") band, so that the electrons which they contain are easily excited up into the conduction band. There, they are free to move about and contribute to the electrical conductivity of the sample. Note also, however, that since the electron has left the vicinity of the phosphorus atom, there is now one more proton than there are electrons at the atom, and hence it has a net positive charge of $1q$. We have represented this by putting a little "+" sign in each P-cup. Note that this positive charge is fixed at the site of the phosphorous atom called a *donor* since it "donates" an electron up into the conduction band, and is not free to move about in the crystal.



Silicon doped with phosphorus.

How many phosphorus atoms do we need to significantly change the resistance of our silicon? Suppose we wanted our 1 mm by 1 mm square sample to have a resistance of one ohm as opposed to more than 60 M Ω .

Turning the resistance equation around we get, [\[link\]](#). And hence, if we continue to assume an electron mobility of 1000 cm²/volt.sec, [\[link\]](#).

Equation:

$$\begin{aligned}\sigma &= L/RA \\ &= 1 \, \Omega/1 \times (0.1)^2 \\ &= 100 \, \text{mho/cm}\end{aligned}$$

Equation:

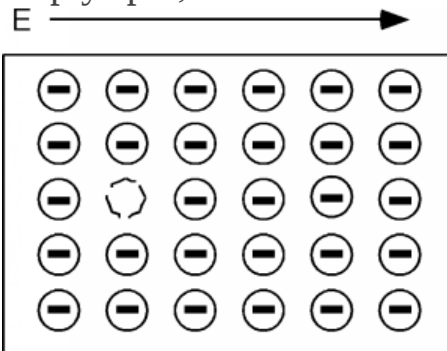
$$\begin{aligned}n &= \sigma/qu \\ &= 100/(1.6 \times 10^{-19})1000 \\ &= 6.25 \times 10^{17} \, \text{cm}^3\end{aligned}$$

Now adding more than 6×10^{17} phosphorus atoms per cubic centimeter might seem like a lot of phosphorus, until you realize that there are almost 10^{24} silicon atoms in a cubic centimeter and hence only one in every 1.6 million silicon atoms has to be changed into a phosphorus one to reduce the resistance of the sample from several 10s of M Ω down to only one Ω . This is the real power of semiconductors. You can make dramatic changes in their electrical properties by the addition of only minute amounts of impurities. This process is called *doping* the semiconductor. It is also one of the great challenges of the semiconductor manufacturing industry, for it is necessary to maintain fantastic levels of control of the impurities in the material in order to predict and control their electrical properties.

Again, if this were the end of the story, we still would not have any calculators, cell phones, or stereos, or at least they would be very big and cumbersome and unreliable, because they would have to work using vacuum tubes. We now have to focus on the few "empty" spots in the lower, almost full band (called the *valence band*.) We will take another view of this band, from a somewhat different perspective. I must confess at this point that what I am giving you is even further from the way things really work, then the "cups at different energies" picture we have been using so far. The problem is, that in order to do things right, we have to get involved in momentum phase-space, a lot more quantum mechanics, and generally a bunch of math and concepts we don't really need in order to have some idea

of how semiconductor devices work. What follow below is really intended as a motivation, so that you will have some feeling that what we state as results, is actually reasonable.

Consider [\[link\]](#). Here we show all of the electrons in the valence, or almost full band, and for simplicity show one missing electron. Let's apply an electric field, as shown by the arrow in the figure. The field will try to move the (negatively charged) electrons to the left, but since the band is almost completely full, the only one that can move is the one right next to the empty spot, or *hole* as it is called.

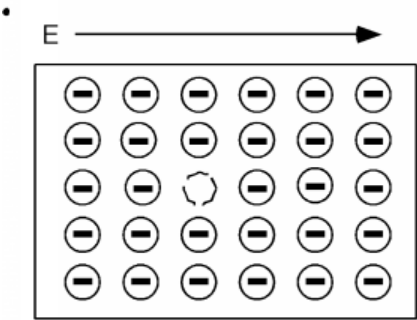


Band full of electrons,
with one missing.

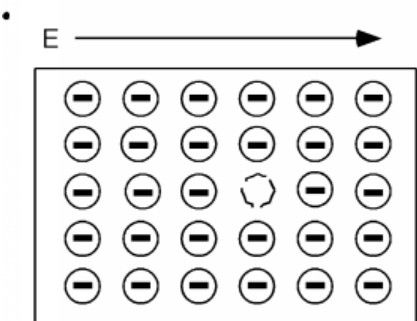
One thing you may be worrying about is what happens to the electrons at the ends of the sample. This is one of the places where we are getting a somewhat distorted view of things, because we should really be looking in momentum, or wave-vector space rather than "real" space. In that picture, they magically drop off one side and "reappear" on the other. This doesn't happen in real space of course, so there is no easy way we can deal with it.

A short time after we apply the electric field we have the situation shown in [\[link\]](#), and a little while after that we have [\[link\]](#). We can interpret this motion in two ways. One is that we have a net flow of negative charge to the left, or if we consider the effect of the aggregate of all the electrons in the band we could picture what is going on as a single positive charge, moving to the right. This is shown in [\[link\]](#). Note that in either view we have the same net effect in the way the total net charge is transported

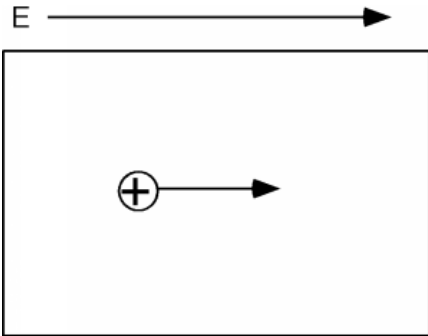
through the sample. In the mostly negative charge picture, we have a net flow of negative charge to the left. In the single positive charge picture, we have a net flow of positive charge to the right. Both give the same sign for the current!



Motion of the
"missing" electron
with an electric field.



Further motion of the
"missing electron"
spot.



Motion of a "hole"
due to an applied
electric field.

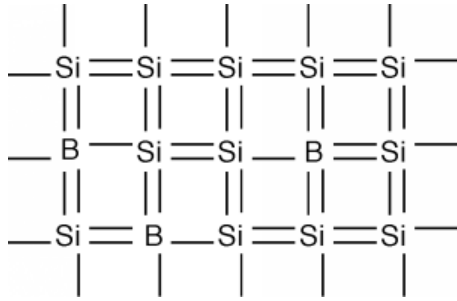
Thus, it turns out, we can consider the consequences of the empty spaces moving through the co-ordinated motion of electrons in an almost full band as being the motion of positive charges, moving wherever these empty spaces happen to be. We call these charge carriers "holes" and they too can add to the total conduction of electricity in a semiconductor. Using ρ to represent the density (in cm^{-3}) of spaces in the valence band and μ_e and μ_h to represent the mobility of electrons and holes respectively (they are usually not the same) we can modify to give the conductivity σ , when both electrons' holes are present, [\[link\]](#).

Equation:

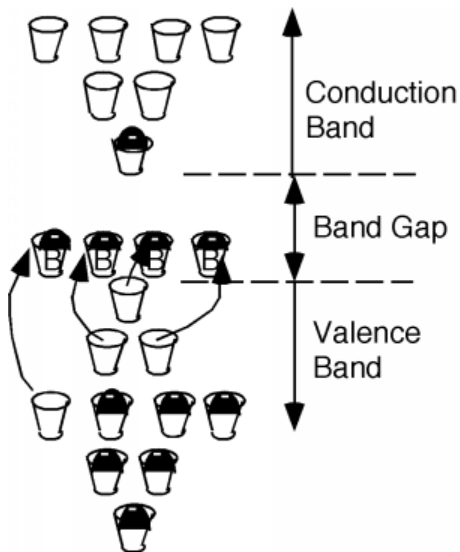
$$\sigma = nq\mu_e + \rho q\mu_h$$

How can we get a sample of semiconductor with a lot of holes in it? What if, instead of phosphorus, we dope our silicon sample with a group III element, say boron? This is shown in [\[link\]](#). Now we have some missing orbitals, or places where electrons could go if they were around. This modifies our energy picture as follows in [\[link\]](#). Now we see a set of new levels introduced by the boron atoms. They are located within the band gap, just a little way above the top of the almost full, or valence band. Electrons in the valence band can be thermally excited up into these new allowed levels, creating empty states, or holes, in the valence band. The excited

electrons are stuck at the boron atom sites called *acceptors*, since they "accept" an electron from the valence band, and hence act as fixed negative charges, localized there. A semiconductor which is doped predominantly with acceptors is called *p-type*, and most of the electrical conduction takes place through the motion of holes. A semiconductor which is doped with donors is called *n-type*, and conduction takes place mainly through the motion of electrons.



A two dimensional representation of a silicon crystal lattice doped with boron.



P-type silicon, due to boron acceptors.

In n-type material, we can assume that all of the phosphorous atoms, or *donors*, are fully ionized when they are present in the silicon structure. Since the number of donors is usually much greater than the native, or intrinsic electron concentration, ($\approx 10^{10} \text{ cm}^{-3}$), if N_d is the density of donors in the material, then n , the electron concentration, $\approx N_d$. If an electron deficient material such as boron is present, then the material is called *p-type* silicon, and the hole concentration is just $\approx N_a$ the concentration of *acceptors*, since these atoms "accept" electrons from the valence band.

If both donors and acceptors are in the material, then whichever one has the higher concentration wins out. This is called compensation. If there are more donors than acceptors then the material is n-type and $n \approx N_d - N_a$. If there are more acceptors than donors then the material is p-type and $p \approx N_a - N_d$. It should be noted that in most compensated material, one type of impurity usually has a much greater (several order of magnitude) concentration than the other, and so the subtraction process described above usually does not change things very much, e.g., $10^{18} - 10^{16} \approx 10^{18}$.

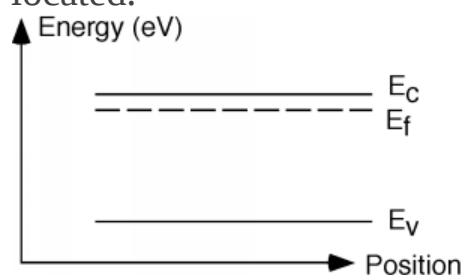
One other fact which you might find useful is that, again, because of quantum mechanics, it turns out that the product of the electron and hole concentration in a material must remain a constant. In silicon at room temperature:

Equation:

$$np \equiv n_i^2 \approx 10^{20} \text{ cm}^{-3}$$

Thus, if we have an n-type sample of silicon doped with 10^{17} donors per cubic centimeter, then n , the electron concentration is just p , the hole concentration, is $10^{20}/10^{17} = 10^3 \text{ cm}^{-3}$. The carriers which dominate a material are called *majority carriers*, which would be the electrons in the above example. The other carriers are called *minority carriers* (the holes in the example) and while 10^3 might not seem like much compared to 10^{17} the presence of minority carriers is still quite important and can not be ignored. Note that if the material is undoped, then it must be that $n = p$ and $n = p = 10^{10}$.

The picture of "cups" of different allowed energy levels is useful for gaining a pictorial understanding of what is going on in a semiconductor, but becomes somewhat awkward when you want to start looking at devices which are made up of both n and p type silicon. Thus, we will introduce one more way of describing what is going on in our material. The picture shown in [\[link\]](#) is called a band diagram. A **band diagram** is just a representation of the energy as a function of position with a semiconductor device. In a band diagram, positive energy for electrons is upward, while for holes, positive energy is downwards. That is, if an electron moves upward, its potential energy increases just as a with a normal mass in a gravitational field. Also, just as a mass will "fall down" if given a chance, an electron will move down a slope shown in a band diagram. On the other hand, holes gain energy by moving downward and so they have a tendency to "float" upward if given the chance - much like a bubble in a liquid. The line labeled E_c in [\[link\]](#) shows the edge of the conduction band, or the bottom of the lowest unoccupied allowed band, while E_v is the top edge of the valence, or highest occupied band. The band gap, E_g for the material is obviously $E_c - E_v$. The dotted line labeled E_f is called the *Fermi level* and it tells us something about the chemical equilibrium energy of the material, and also something about the type and number of carriers in the material. More on this later. Note that there is no zero energy level on a diagram such as this. We often use either the Fermi level or one or other of the band edges as a reference level on lieu of knowing exactly where "zero energy" is located.



An electron band-
diagram for a
semiconductor.

The distance (in energy) between the Fermi level and either E_c and E_v gives us information concerning the density of electrons and holes in that region of the semiconductor material. The details, once again, will have to be begged off on grounds of mathematical complexity. It turns out that you can say:

Equation:

$$n = N_c e^{-\left(\frac{E_c - E_f}{kT}\right)}$$

Equation:

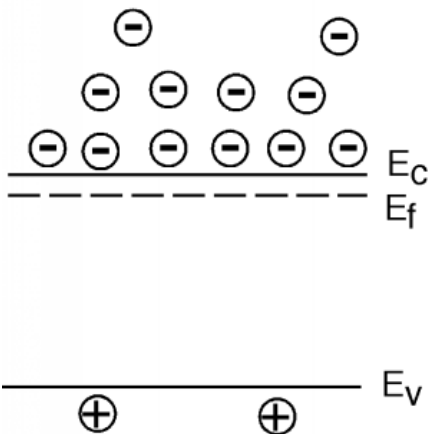
$$p = N_v e^{-\left(\frac{E_f - E_v}{kT}\right)}$$

Both N_c and N_v are constants that depend on the material you are talking about, but are typically on the order of 10^{19} cm^{-3} . The expression in the denominator of the exponential is just Boltzman's constant ($8.63 \times 10^{-5} \text{ eV/K}$), k , times the temperature T of the material (in absolute temperature or Kelvin). At room temperature $kT = 1/40$ of an electron volt. Look carefully at the numerators in the exponential. Note first that there is a minus sign in front, which means the bigger the number in the exponent, the fewer carriers we have. Thus, the top expression says that if we have n-type material, then E_f must not be too far away from the conduction band, while if we have p-type material, then the Fermi level, E_f must be down close to the valence band. The closer E_f gets to E_c the more electrons we have. The closer E_f gets to E_v , the more holes we have. [\[link\]](#) therefore must be for a sample of n-type material. Note also that if we know how heavily a sample is doped (i.e., we know what N_d is) and from the fact that $n \approx N_d$ we can use [\[link\]](#) to find out how far away the Fermi level is from the conduction band, [\[link\]](#).

Equation:

$$E_f - E_c = kT \ln\left(\frac{N_c}{N_d}\right)$$

To help further in our ability to picture what is going on, we will often add to this band diagram, some small signed circles to indicate the presence of mobile electrons and holes in the material. Note that the electrons are spread out in energy. From our "cups" picture we know they like to stay in the lower energy states if possible, but some will be distributed into the higher levels as well. What is distorted here is the scale. The band-gap for silicon is 1.1 eV, while the actual spread of the electrons would probably only be a few tenths of an eV, not nearly as much as is shown in [\[link\]](#). Lets look at a sample of p-type material, just for comparison. Note that for holes, increasing energy goes *down* not up, so their distribution is inverted from that of the electrons. You can kind of think of holes as bubbles in a glass of soda or beer, they want to float to the top if they can. Note also for both n and p-type material there are also a few "minority" carriers, or carriers of the opposite type, which arise from thermal generation across the band-gap.



Band diagram for an
n-type
semiconductor.

Diffusion

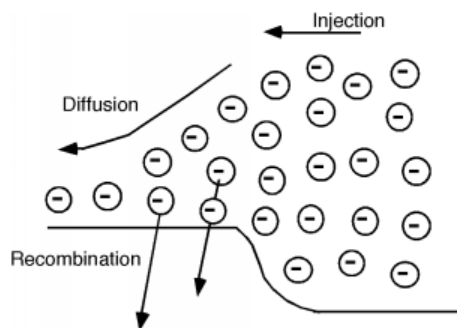
The module discusses the process of electrons moving across a p-n or n-p junction known as diffusion.

Note: This module is adapted from the Connexions module entitled *Diffusion* by Bill Wilson.

Introduction

Let us turn our attention to what happens to the electrons and holes once they have been injected across a forward-biased junction. We will concentrate just on the electrons which are injected into the p-side of the junction, but keep in mind that similar things are also happening to the holes which enter the n-side.

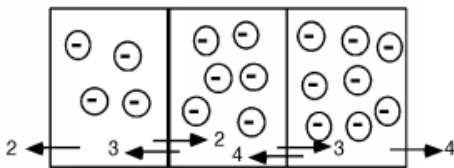
When electrons are injected across a junction, they move away from the junction region by a diffusion process, while at the same time, some of them are disappearing because they are minority carriers (electrons in basically p-type material) and so there are lots of holes around for them to recombine with. This is all shown schematically in [\[link\]](#).



Processes involved in
electron transport
across a p-n junction.

Diffusion process quantified

It is actually fairly easy to quantify this, and come up with an expression for the electron distribution within the p-region. First we have to look a little bit at the diffusion process however. Imagine that we have a series of bins, each with a different number of electrons in them. In a given time, we could imagine that all of the electrons would flow out of their bins into the neighboring ones. Since there is no reason to expect the electrons to favor one side over the other, we will assume that exactly half leave by each side. This is all shown in [\[link\]](#). We will keep things simple and only look at three bins. Imagine there are 4, 6, and 8 electrons respectively in each of the bins. After the required "emptying time," we will have a net flux of exactly one electron across each boundary as shown.



A schematic representation of a diffusion problem.

Now let's raise the number of electrons to 8, 12 and 16 respectively ([\[link\]](#)). We find that the net flux across each boundary is now 2 electrons per emptying time, rather than one. Note that the gradient (slope) of the concentration in the boxes has also doubled from one per box to two per box. This leads us to a rather obvious statement that the flux of carriers is proportional to the gradient of their density. This is stated formally in what

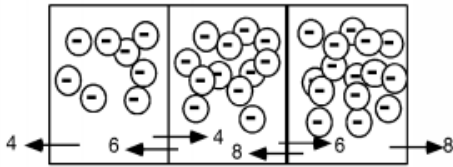
is known as Fick's First Law of Diffusion, [\[link\]](#). Where D_e is simply a proportionality constant called the diffusion coefficient. Since we are talking about the motion of electrons, this diffusion flux must give rise to a current density $J_{e\text{diff}}$. Since an electron has a charge $-q$ associated with it, [\[link\]](#).

Equation:

$$\text{Flux} = (-D_e) \frac{d n(x)}{d x}$$

Equation:

$$J_{e\text{diff}} = qD_e \frac{d n}{d x}$$



A schematic
representation of a
diffusion from bins.

Now we have to invoke something called the continuity equation. Imagine we have a volume (V) which is filled with some charge (Q). It is fairly obvious that if we add up all of the current density which is flowing out of the volume that it must be equal to the time rate of decrease of the charge within that volume. This idea is expressed in the formula below which uses a closed-surface integral, along with the all the other integrals to follow:

Equation:

$$\oint_S J \, dS = - \frac{dQ}{dt}$$

We can write Q as, [\[link\]](#), where we are doing a volume integral of the charge density (ρ) over the volume (V). Now we can use Gauss' theorem which says we can replace a surface integral of a quantity with a volume integral of its divergence, [\[link\]](#).

Equation:

$$Q = \int_V \rho(v) \, dV$$

Equation:

$$\oint_S J \, dS = \int \operatorname{div} (J) \, dV$$

So, combining [\[link\]](#), [\[link\]](#) and [\[link\]](#), we have, [\[link\]](#).

Equation:

$$\int \operatorname{div} (J) \, dV = - \int \frac{d\rho}{dt} \, dV$$

Finally, we let the volume V shrink down to a point, which means the quantities inside the integral must be equal, and we have the differential form of the continuity equation (in one dimension), [\[link\]](#).

Equation:

$$\begin{aligned} \operatorname{div} (J) &= \frac{\partial J}{\partial x} \\ &= - \frac{d\rho(x)}{dt} \end{aligned}$$

What about the electrons?

Now let's go back to the electrons in the diode. The electrons which have been injected across the junction are called *excess minority carriers*, because they are electrons in a p-region (hence minority) but their concentration is greater than what they would be if they were in a sample of p-type material at equilibrium. We will designate them as n' , and since they could change with both time and position we shall write them as $n'(x,t)$. Now there are two ways in which $n'(x,t)$ can change with time. One would be if we were to stop injecting electrons in from the n-side of the junction. A reasonable way to account for the decay which would occur if we were not supplying electrons would be to write:

Equation:

$$\frac{d}{dt} n'(x, t) = - \frac{n'(x, t)}{\tau_r}$$

Where τ_r called the minority carrier recombination lifetime. It is pretty easy to show that if we start out with an excess minority carrier concentration n'_0 at $t = 0$, then $n'(x,t)$ will go as, [\[link\]](#). But, the electron concentration can also change because of electrons flowing into or out of the region x . The electron concentration $n'(x,t)$ is just $\frac{\rho(x,t)}{q}$. Thus, due to electron flow we have, [\[link\]](#).

Equation:

$$n'(x, t) = n'_0 e^{\frac{-t}{\tau_r}}$$

Equation:

$$\begin{aligned} \frac{d}{dt} n'(x, t) &= \frac{1}{q} \frac{d\rho(x,t)}{dt} \\ &= \frac{1}{q} \operatorname{div} (J(x, t)) \end{aligned}$$

But, we can get an expression for $J(x, t)$ from [\[link\]](#). Reducing the divergence in [\[link\]](#) to one dimension (we just have a $\frac{\partial J}{\partial x}$) we finally end up

with, [\[link\]](#).

Equation:

$$\frac{d}{dt} n'(x, t) = D_e \frac{d^2 n'(x, t)}{dx^2}$$

Combining [\[link\]](#) and [\[link\]](#) (electrons will, after all, suffer from both recombination and diffusion) and we end up with:

Equation:

$$\frac{d}{dt} n'(x, t) = D_e \frac{d^2 n'(x, t)}{dx^2} - \frac{n'(x, t)}{\tau_r}$$

This is a somewhat specialized form of an equation called the ambipolar diffusion equation. It seems kind of complicated but we can get some nice results from it if we make some simply boundary condition assumptions.

Using the ambipolar diffusion equation

For anything we will be interested in, we will only look at steady state solutions. This means that the time derivative on the LHS of [\[link\]](#) is zero, and so letting $n'(x, t)$ become simply $n'(x)$ since we no longer have any time variation to worry about, we have:

Equation:

$$\frac{d^2}{dx^2} n'(x) - \frac{1}{D_e \tau_r} n'(x) = 0$$

Picking the not unreasonable boundary conditions that $n'(0) = n_0$ (the concentration of excess electrons just at the start of the diffusion region) and $n'(x) \rightarrow 0$ as $x \rightarrow \infty$ (the excess carriers go to zero when we get far from the junction) then:

Equation:

$$n(x) = n_0 e^{-\frac{x}{\sqrt{D_e \tau_r}}}$$

The expression in the radical $\sqrt{D_e \tau_r}$ is called the electron diffusion length, L_e , and gives us some idea as to how far away from the junction the excess electrons will exist before they have more or less all recombined. This will be important for us when we move on to bipolar transistors. A typical value for the diffusion coefficient for electrons in silicon would be $D_e = 25 \text{ cm}^2/\text{sec}$ and the minority carrier lifetime is usually around a microsecond. As shown in [\[link\]](#) this is not very far at all.

Equation:

$$\begin{aligned} L_e &= \sqrt{D_e \tau_r} \\ &= \sqrt{25 \times 10^{-6}} \\ &= 5 \times 10^{-3} \text{ cm} \end{aligned}$$

Crystal Structure

Introduction

In any sort of discussion of crystalline materials, it is useful to begin with a discussion of crystallography: the study of the formation, structure, and properties of crystals. A crystal structure is defined as the particular repeating arrangement of atoms (molecules or ions) throughout a crystal. Structure refers to the internal arrangement of particles and not the external appearance of the crystal. However, these are not entirely independent since the external appearance of a crystal is often related to the internal arrangement. For example, crystals of cubic rock salt (NaCl) are physically cubic in appearance. Only a few of the possible crystal structures are of concern with respect to simple inorganic salts and these will be discussed in detail, however, it is important to understand the nomenclature of crystallography.

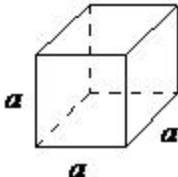
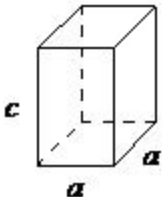
Crystallography

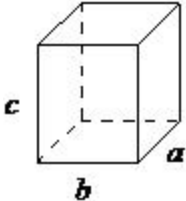
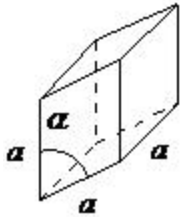
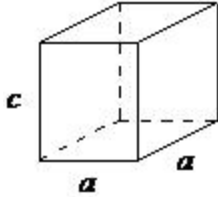
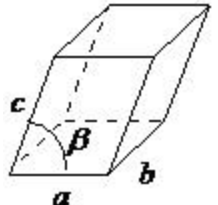
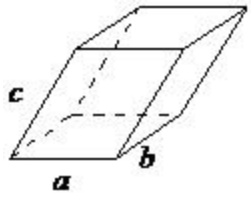
Bravais lattice

The Bravais lattice is the basic building block from which all crystals can be constructed. The concept originated as a topological problem of finding the number of different ways to arrange points in space where each point would have an identical “atmosphere”. That is each point would be surrounded by an identical set of points as any other point, so that all points would be indistinguishable from each other. Mathematician Auguste Bravais discovered that there were 14 different collections of the groups of points, which are known as Bravais lattices. These lattices fall into seven different “crystal systems”, as differentiated by the relationship between the angles between sides of the “unit cell” and the distance between points in the unit cell. The unit cell is the smallest group of atoms, ions or molecules that, when repeated at regular intervals in three dimensions, will produce the lattice of a crystal system. The “lattice parameter” is the length between two points on the corners of a unit cell. Each of the various lattice parameters are designated by the letters a , b , and c . If two sides are equal,

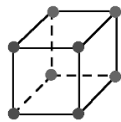
such as in a tetragonal lattice, then the lengths of the two lattice parameters are designated a and c , with b omitted. The angles are designated by the Greek letters α , β , and γ , such that an angle with a specific Greek letter is not subtended by the axis with its Roman equivalent. For example, α is the included angle between the b and c axis.

[\[link\]](#) shows the various crystal systems, while [\[link\]](#) shows the 14 Bravais lattices. It is important to distinguish the characteristics of each of the individual systems. An example of a material that takes on each of the Bravais lattices is shown in [\[link\]](#).

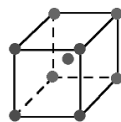
System	Axial lengths and angles	Unit cell geometry
cubic	$a = b = c, \alpha = \beta = \gamma = 90^\circ$	
tetragonal	$a = b \neq c, \alpha = \beta = \gamma = 90^\circ$	
orthorhombic	$a \neq b \neq c, \alpha = \beta = \gamma = 90^\circ$	

		
rhombohedral	$a = b = c, \alpha = \beta = \gamma \neq 90^\circ$	
hexagonal	$a = b \neq c, \alpha = \beta = 90^\circ, \gamma = 120^\circ$	
monoclinic	$a \neq b \neq c, \alpha = \gamma = 90^\circ, \beta \neq 90^\circ$	
triclinic	$a \neq b \neq c, \alpha \neq \beta \neq \gamma$	

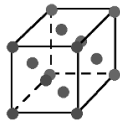
Geometrical characteristics of the seven crystal systems.



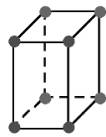
simple cubic



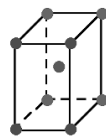
body-centered
cubic



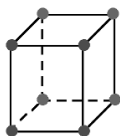
face-centered
cubic



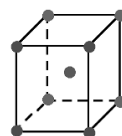
simple
tetragonal



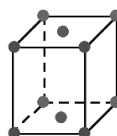
body-centered
tetragonal



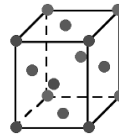
simple
orthorhombic



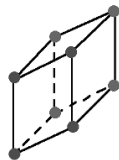
body-centered
orthorhombic



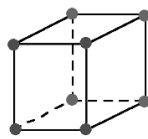
base-centered
orthorhombic



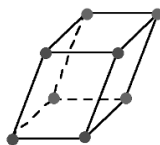
face-centered
orthorhombic



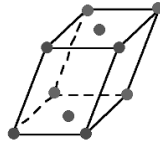
rhombohedral



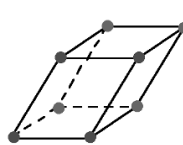
hexagonal



simple
monoclinic



base-centered
monoclinic



triclinic

Bravais lattices.

Crystal system	Example

triclinic	$\text{K}_2\text{S}_2\text{O}_8$
monoclinic	As_4S_4 , KNO_2
rhombohedral	Hg, Sb
hexagonal	Zn, Co, NiAs
orthorhombic	Ga, Fe_3C
tetragonal	In, TiO_2
cubic	Au, Si, NaCl

Examples of elements and compounds that adopt each of the crystal systems.

The cubic lattice is the most symmetrical of the systems. All the angles are equal to 90° , and all the sides are of the same length ($a = b = c$). Only the length of one of the sides (a) is required to describe this system completely. In addition to simple cubic, the cubic lattice also includes body-centered cubic and face-centered cubic ([\[link\]](#)). Body-centered cubic results from the presence of an atom (or ion) in the center of a cube, in addition to the atoms (ions) positioned at the vertices of the cube. In a similar manner, a face-centered cubic requires, in addition to the atoms (ions) positioned at the vertices of the cube, the presence of atoms (ions) in the center of each of the cubes face.

The tetragonal lattice has all of its angles equal to 90° , and has two out of the three sides of equal length ($a = b$). The system also includes body-centered tetragonal ([\[link\]](#)).

In an orthorhombic lattice all of the angles are equal to 90° , while all of its sides are of unequal length. The system needs only to be described by three lattice parameters. This system also includes body-centered orthorhombic, base-centered orthorhombic, and face-centered orthorhombic ([\[link\]](#)). A base-centered lattice has, in addition to the atoms (ions) positioned at the

vertices of the orthorhombic lattice, atoms (ions) positioned on just two opposing faces.

The rhombohedral lattice is also known as trigonal, and has no angles equal to 90° , but all sides are of equal length ($a = b = c$), thus requiring only by one lattice parameter, and all three angles are equal ($\alpha = \beta = \gamma$).

A hexagonal crystal structure has two angles equal to 90° , with the other angle (γ) equal to 120° . For this to happen, the two sides surrounding the 120° angle must be equal ($a = b$), while the third side (c) is at 90° to the other sides and can be of any length.

The monoclinic lattice has no sides of equal length, but two of the angles are equal to 90° , with the other angle (usually defined as β) being something other than 90° . It is a tilted parallelogram prism with rectangular bases. This system also includes base-centered monoclinic ([\[link\]](#)).

In the triclinic lattice none of the sides of the unit cell are equal, and none of the angles within the unit cell are equal to 90° . The triclinic lattice is chosen such that all the internal angles are either acute or obtuse. This crystal system has the lowest symmetry and must be described by 3 lattice parameters (a , b , and c) and the 3 angles (α , β , and γ).

Atom positions, crystal directions and Miller indices

Atom positions and crystal axes

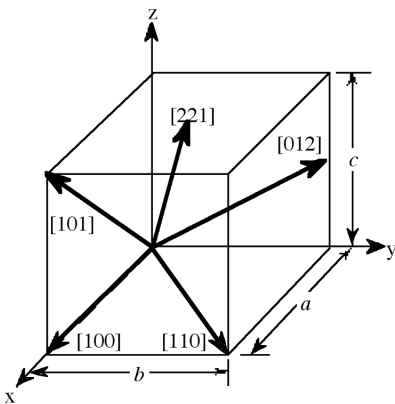
The structure of a crystal is defined with respect to a unit cell. As the entire crystal consists of repeating unit cells, this definition is sufficient to represent the entire crystal. Within the unit cell, the atomic arrangement is expressed using coordinates. There are two systems of coordinates commonly in use, which can cause some confusion. Both use a corner of the unit cell as their origin. The first, less-commonly seen system is that of Cartesian or orthogonal coordinates (X , Y , Z). These usually have the units of Angstroms and relate to the distance in each direction between the origin

of the cell and the atom. These coordinates may be manipulated in the same fashion are used with two- or three-dimensional graphs. It is very simple, therefore, to calculate inter-atomic distances and angles given the Cartesian coordinates of the atoms. Unfortunately, the repeating nature of a crystal cannot be expressed easily using such coordinates. For example, consider a cubic cell of dimension 3.52 Å. Pretend that this cell contains an atom that has the coordinates (1.5, 2.1, 2.4). That is, the atom is 1.5 Å away from the origin in the x direction (which coincides with the *a* cell axis), 2.1 Å in the y (which coincides with the *b* cell axis) and 2.4 Å in the z (which coincides with the *c* cell axis). There will be an equivalent atom in the next unit cell along the x-direction, which will have the coordinates (1.5 + 3.52, 2.1, 2.4) or (5.02, 2.1, 2.4). This was a rather simple calculation, as the cell has very high symmetry and so the cell axes, *a*, *b* and *c*, coincide with the Cartesian axes, X, Y and Z. However, consider lower symmetry cells such as triclinic or monoclinic in which the cell axes are not mutually orthogonal. In such cases, expressing the repeating nature of the crystal is much more difficult to accomplish.

Accordingly, atomic coordinates are usually expressed in terms of fractional coordinates, (x, y, z). This coordinate system is coincident with the cell axes (*a*, *b*, *c*) and relates to the position of the atom in terms of the fraction along each axis. Consider the atom in the cubic cell discussion above. The atom was 1.5 Å in the *a* direction away from the origin. As the *a* axis is 3.52 Å long, the atom is $(1.5/3.52)$ or 0.43 of the axis away from the origin. Similarly, it is $(2.1/3.52)$ or 0.60 of the *b* axis and $(2.4/3.52)$ or 0.68 of the *c* axis. The fractional coordinates of this atom are, therefore, (0.43, 0.60, 0.68). The coordinates of the equivalent atom in the next cell over in the *a* direction, however, are easily calculated as this atom is simply 1 unit cell away in *a*. Thus, all one has to do is add 1 to the x coordinate: (1.43, 0.60, 0.68). Such transformations can be performed regardless of the shape of the unit cell. Fractional coordinates, therefore, are used to retain and manipulate crystal information.

Crystal directions

The designation of the individual vectors within any given crystal lattice is accomplished by the use of whole number multipliers of the lattice parameter of the point at which the vector exits the unit cell. The vector is indicated by the notation $[hkl]$, where h , k , and l are reciprocals of the point at which the vector exits the unit cell. The origination of all vectors is assumed defined as $[000]$. For example, the direction along the a -axis according to this scheme would be $[100]$ because this has a component only in the a -direction and no component along either the b or c axial direction. A vector diagonally along the face defined by the a and b axis would be $[110]$, while going from one corner of the unit cell to the opposite corner would be in the $[111]$ direction. [\[link\]](#) shows some examples of the various directions in the unit cell. The crystal direction notation is made up of the lowest combination of integers and represents unit distances rather than actual distances. A $[222]$ direction is identical to a $[111]$, so $[111]$ is used. Fractions are not used. For example, a vector that intercepts the center of the top face of the unit cell has the coordinates $x = 1/2$, $y = 1/2$, $z = 1$. All have to be inversed to convert to the lowest combination of integers (whole numbers); i.e., $[221]$ in [\[link\]](#). Finally, all parallel vectors have the same crystal direction, e.g., the four vertical edges of the cell shown in [\[link\]](#) all have the crystal direction $[hkl] = [001]$.



Some common directions in a cubic unit cell.

Crystal directions may be grouped in families. To avoid confusion there exists a convention in the choice of brackets surrounding the three numbers to differentiate a crystal direction from a family of direction. For a direction, square brackets $[hkl]$ are used to indicate an individual direction. Angle brackets $\langle hkl \rangle$ indicate a family of directions. A family of directions includes any directions that are equivalent in length and types of atoms encountered. For example, in a cubic lattice, the $[100]$, $[010]$, and $[001]$ directions all belong to the $\langle 100 \rangle$ family of planes because they are equivalent. If the cubic lattice were rotated 90° , the a , b , and c directions would remain indistinguishable, and there would be no way of telling on which crystallographic positions the atoms are situated, so the family of directions is the same. In a hexagonal crystal, however, this is not the case, so the $[100]$ and $[010]$ would both be $\langle 100 \rangle$ directions, but the $[001]$ direction would be distinct. Finally, negative directions are identified with a bar over the negative number instead of a minus sign.

Crystal planes

Planes in a crystal can be specified using a notation called Miller indices. The Miller index is indicated by the notation $[hkl]$ where h , k , and l are reciprocals of the plane with the x , y , and z axes. To obtain the Miller indices of a given plane requires the following steps:

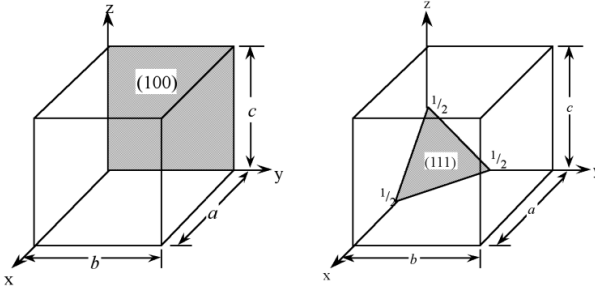
The plane in question is placed on a unit cell.

Its intercepts with each of the crystal axes are then found.

The reciprocal of the intercepts are taken.

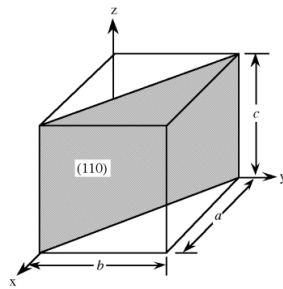
These are multiplied by a scalar to insure that is in the simple ratio of whole numbers.

For example, the face of a lattice that does not intersect the y or z axis would be (100) , while a plane along the body diagonal would be the (111) plane. An illustration of this along with the (111) and (110) planes is given in [\[link\]](#).



$$\begin{matrix} h & k & l \\ \frac{1}{1}, \frac{1}{\infty}, \frac{1}{\infty} \end{matrix} = (100)$$

$$\begin{matrix} h & k & l \\ \frac{1}{1/2}, \frac{1}{1/2}, \frac{1}{1/2} \end{matrix} = (222) = (111)$$



$$\begin{matrix} h & k & l \\ \frac{1}{1}, \frac{1}{1}, \frac{1}{\infty} \end{matrix} = (110)$$

Examples of Miller indices notation for crystal planes.

As with crystal directions, Miller indices directions may be grouped in families. Individual Miller indices are given in parentheses (hkl), while braces $\{hkl\}$ are placed around the indices of a family of planes. For example, (001), (100), and (010) are all in the $\{100\}$ family of planes, for a cubic lattice.

Description of crystal structures

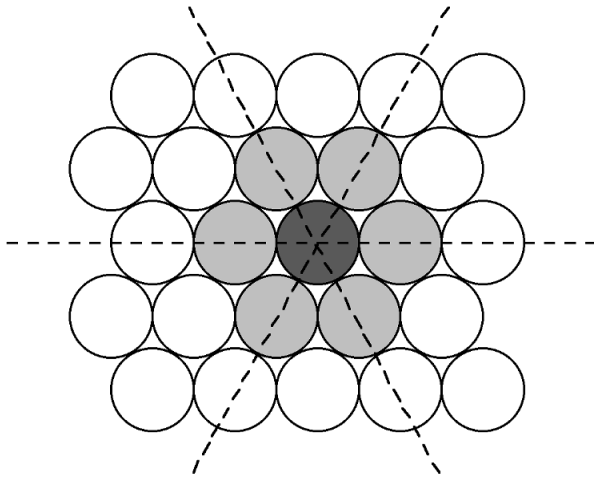
Crystal structures may be described in a number of ways. The most common manner is to refer to the size and shape of the unit cell and the positions of the atoms (or ions) within the cell. However, this information is sometimes insufficient to allow for an understanding of the true structure in three dimensions. Consideration of several unit cells, the arrangement of the

atoms with respect to each other, the number of other atoms they in contact with, and the distances to neighboring atoms, often will provide a better understanding. A number of methods are available to describe extended solid-state structures. The most applicable with regard to elemental and compound semiconductor, metals and the majority of insulators is the close packing approach.

Close packed structures: hexagonal close packing and cubic close packing

Many crystal structures can be described using the concept of close packing. This concept requires that the atoms (ions) are arranged so as to have the maximum density. In order to understand close packing in three dimensions, the most efficient way for equal sized spheres to be packed in two dimensions must be considered.

The most efficient way for equal sized spheres to be packed in two dimensions is shown in [\[link\]](#), in which it can be seen that each sphere (the dark gray shaded sphere) is surrounded by, and is in contact with, six other spheres (the light gray spheres in [\[link\]](#)). It should be noted that contact with six other spheres the maximum possible is the spheres are the same size, although lower density packing is possible. Close packed layers are formed by repetition to an infinite sheet. Within these close packed layers, three close packed rows are present, shown by the dashed lines in [\[link\]](#).

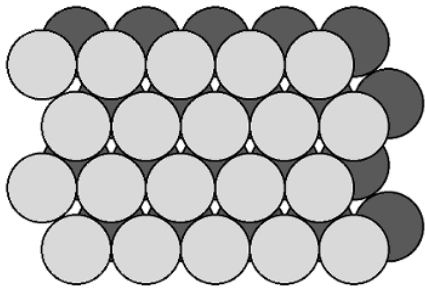


Schematic representation of a close packed layer of equal sized spheres. The close packed rows (directions) are shown by the dashed lines.

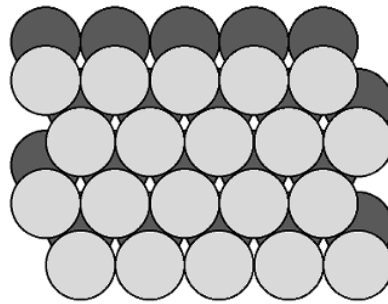
The most efficient way for equal sized spheres to be packed in three dimensions is to stack close packed layers on top of each other to give a close packed structure. There are two simple ways in which this can be done, resulting in either a hexagonal or cubic close packed structures.

Hexagonal close packed

If two close packed layers A and B are placed in contact with each other so as to maximize the density, then the spheres of layer B will rest in the hollow (vacancy) between three of the spheres in layer A. This is demonstrated in [\[link\]](#). Atoms in the second layer, B (shaded light gray), may occupy one of two possible positions ([\[link\]](#)a or b) but not both together or a mixture of each. If a third layer is placed on top of layer B such that it exactly covers layer A, subsequent placement of layers will result in the following sequence ...ABABAB.... This is known as hexagonal close packing or *hcp*.



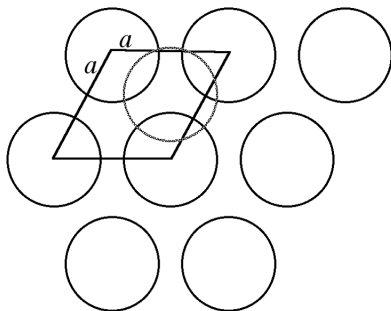
(a)



(b)

Schematic representation of two close packed layers arranged in A (dark grey) and B (light grey) positions. The alternative stacking of the B layer is shown in (a) and (b).

The hexagonal close packed cell is a derivative of the hexagonal Bravais lattice system ([\[link\]](#)) with the addition of an atom inside the unit cell at the coordinates $(\frac{1}{3}, \frac{2}{3}, \frac{1}{2})$. The basal plane of the unit cell coincides with the close packed layers ([\[link\]](#)). In other words the close packed layer makes-up the $\{001\}$ family of crystal planes.

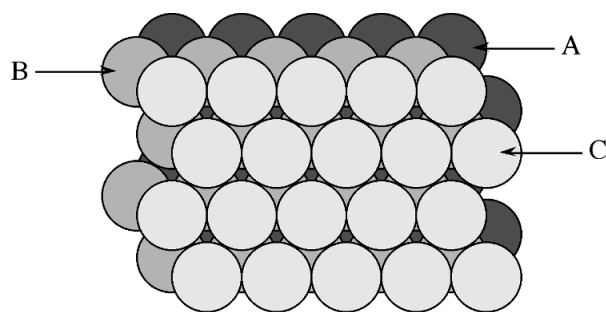


A schematic projection of the basal plane of the hcp unit cell on the close packed layers.

The “packing fraction” in a hexagonal close packed cell is 74.05%; that is 74.05% of the total volume is occupied. The packing fraction or density is derived by assuming that each atom is a hard sphere in contact with its nearest neighbors. Determination of the packing fraction is accomplished by calculating the number of whole spheres per unit cell (2 in hcp), the volume occupied by these spheres, and a comparison with the total volume of a unit cell. The number gives an idea of how “open” or filled a structure is. By comparison, the packing fraction for body-centered cubic ([link](#)) is 68% and for diamond cubic (an important semiconductor structure to be described later) is it 34%.

Cubic close packed: face-centered cubic

In a similar manner to the generation of the hexagonal close packed structure, two close packed layers are stacked ([link](#)) however, the third layer (C) is placed such that it does not exactly cover layer A, while sitting in a set of troughs in layer B ([link](#)), then upon repetition the packing sequence will be ...ABCABCABC.... This is known as cubic close packing or *ccp*.



Schematic representation of the three close packed layers in a cubic close packed arrangement: A (dark grey), B

(medium grey), and C (light grey).

The unit cell of cubic close packed structure is actually that of a face-centered cubic (*fcc*) Bravais lattice. In the *fcc* lattice the close packed layers constitute the {111} planes. As with the *hcp* lattice packing fraction in a cubic close packed (*fcc*) cell is 74.05%. Since face centered cubic or *fcc* is more commonly used in preference to cubic close packed (*ccp*) in describing the structures, the former will be used throughout this text.

Coordination number

The coordination number of an atom or ion within an extended structure is defined as the number of nearest neighbor atoms (ions of opposite charge) that are in contact with it. A slightly different definition is often used for atoms within individual molecules: the number of donor atoms associated with the central atom or ion. However, this distinction is rather artificial, and both can be employed.

The coordination numbers for metal atoms in a molecule or complex are commonly 4, 5, and 6, but all values from 2 to 9 are known and a few examples of higher coordination numbers have been reported. In contrast, common coordination numbers in the solid state are 3, 4, 6, 8, and 12. For example, the atom in the center of body-centered cubic lattice has a coordination number of 8, because it touches the eight atoms at the corners of the unit cell, while an atom in a simple cubic structure would have a coordination number of 6. In both *fcc* and *hcp* lattices each of the atoms have a coordination number of 12.

Octahedral and tetrahedral vacancies

As was mentioned above, the packing fraction in both *fcc* and *hcp* cells is 74.05%, leaving 25.95% of the volume unfilled. The unfilled lattice sites (interstices) between the atoms in a cell are called interstitial sites or vacancies. The shape and relative size of these sites is important in controlling the position of additional atoms. In both *fcc* and *hcp* cells most of the space within these atoms lies within two different sites known as octahedral sites and tetrahedral sites. The difference between the two lies in their “coordination number”, or the number of atoms surrounding each site. Tetrahedral sites (vacancies) are surrounded by four atoms arranged at the corners of a tetrahedron. Similarly, octahedral sites are surrounded by six atoms which make-up the apices of an octahedron. For a given close packed lattice an octahedral vacancy will be larger than a tetrahedral vacancy.

Within a face centered cubic lattice, the eight tetrahedral sites are positioned within the cell, at the general fractional coordinate of $(\frac{n}{4}, \frac{n}{4}, \frac{n}{4})$ where $n = 1$ or 3 , e.g., $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $(\frac{1}{4}, \frac{1}{4}, \frac{3}{4})$, etc. The octahedral sites are located at the center of the unit cell $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, as well as at each of the edges of the cell, e.g., $(\frac{1}{2}, 0, 0)$. In the hexagonal close packed system, the tetrahedral sites are at $(0, 0, \frac{3}{8})$ and $(\frac{1}{3}, \frac{2}{3}, \frac{7}{8})$, and the octahedral sites are at $(\frac{1}{3}, \frac{1}{3}, \frac{1}{4})$ and all symmetry equivalent positions.

Important structure types

The majority of crystalline materials do not have a structure that fits into the one atom per site simple Bravais lattice. A number of other important crystal structures are found, however, only a few of these crystal structures are those of which occur for the elemental and compound semiconductors and the majority of these are derived from *fcc* or *hcp* lattices. Each structural type is generally defined by an archetype, a material (often a naturally occurring mineral) which has the structure in question and to which all the similar materials are related. With regard to commonly used elemental and compound semiconductors the important structures are diamond, zinc blende, Wurtzite, and to a lesser extent chalcopyrite. However, rock salt, β -tin, cinnabar and cesium chloride are observed as high pressure or high temperature phases and are therefore also discussed.

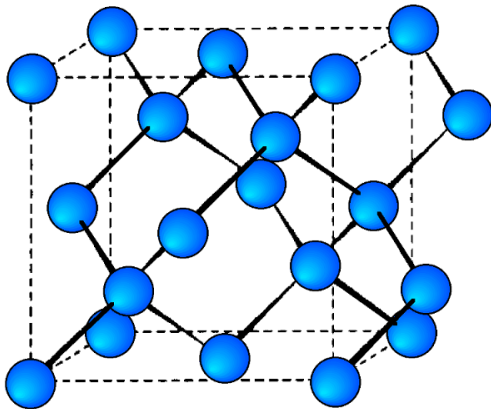
The following provides a summary of these structures. Details of the full range of solid-state structures are given elsewhere.

Diamond Cubic

The diamond cubic structure consists of two interpenetrating face-centered cubic lattices, with one offset $\frac{1}{4}$ of a cube along the cube diagonal. It may also be described as face centered cubic lattice in which half of the tetrahedral sites are filled while all the octahedral sites remain vacant. The diamond cubic unit cell is shown in [\[link\]](#). Each of the atoms (e.g., C) is four coordinate, and the shortest interatomic distance (C-C) may be determined from the unit cell parameter (a).

Equation:

$$\text{C-C} = a \frac{\sqrt{3}}{4} \approx 0.422 a$$



Unit cell structure of a diamond cubic lattice showing the two interpenetrating face-centered cubic lattices.

Zinc blende

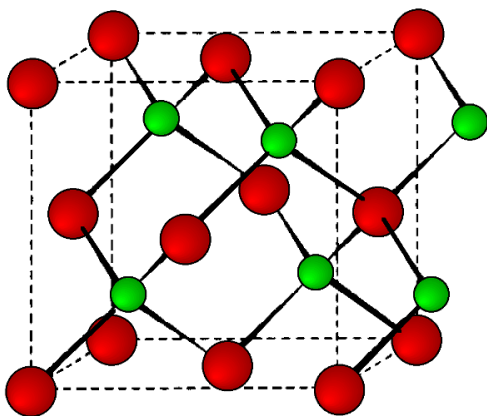
This is a binary phase (ME) and is named after its archetype, a common mineral form of zinc sulfide (ZnS). As with the diamond lattice, zinc blende consists of the two interpenetrating *fcc* lattices. However, in zinc blende one lattice consists of one of the types of atoms (Zn in ZnS), and the other lattice is of the second type of atom (S in ZnS). It may also be described as face centered cubic lattice of S atoms in which half of the tetrahedral sites are filled with Zn atoms. All the atoms in a zinc blende structure are 4-coordinate. The zinc blende unit cell is shown in [\[link\]](#). A number of inter-atomic distances may be calculated for any material with a zinc blende unit cell using the lattice parameter (a).

Equation:

$$\text{Zn-S} = a \frac{\sqrt{3}}{4} \approx 0.422 a$$

Equation:

$$\text{Zn-Zn} = \text{S-S} = \frac{a}{\sqrt{2}} \approx 0.707 a$$

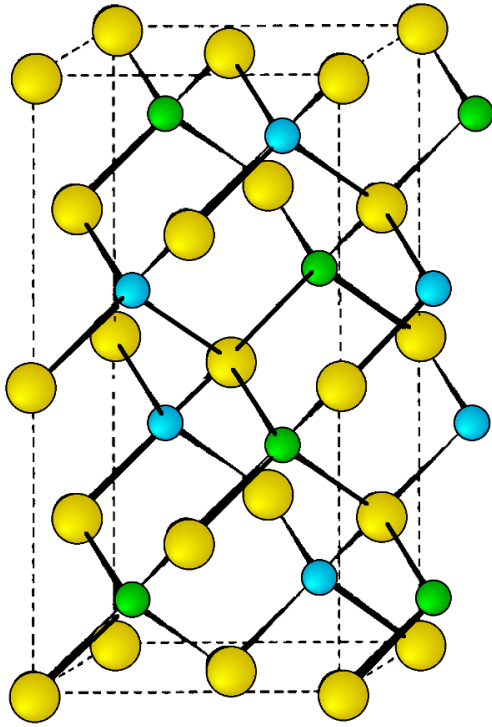


Unit cell structure of a

zinc blende (ZnS) lattice.
Zinc atoms are shown in
green (small), sulfur
atoms shown in red
(large), and the dashed
lines show the unit cell.

Chalcopyrite

The mineral chalcopyrite CuFeS_2 is the archetype of this structure. The structure is tetragonal ($a = b \neq c$, $\alpha = \beta = \gamma = 90^\circ$), and is essentially a superlattice on that of zinc blende. Thus, it is easiest to imagine that the chalcopyrite lattice is made-up of a lattice of sulfur atoms in which the tetrahedral sites are filled in layers, ...FeCuCuFe..., etc. ([\[link\]](#)). In such an idealized structure $c = 2a$, however, this is not true of all materials with chalcopyrite structures.



Unit cell structure of a chalcopyrite lattice. Copper atoms are shown in blue, iron atoms are shown in green and sulfur atoms are shown in yellow. The dashed lines show the unit cell.

Rock salt

As its name implies the archetypal rock salt structure is NaCl (table salt). In common with the zinc blende structure, rock salt consists of two interpenetrating face-centered cubic lattices. However, the second lattice is offset $1/2a$ along the unit cell axis. It may also be described as face centered cubic lattice in which all of the octahedral sites are filled, while all the tetrahedral sites remain vacant, and thus each of the atoms in the rock salt

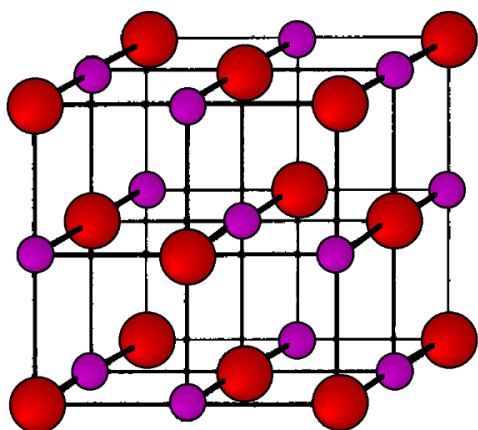
structure are 6-coordinate. The rock salt unit cell is shown in [\[link\]](#). A number of inter-atomic distances may be calculated for any material with a rock salt structure using the lattice parameter (a).

Equation:

$$\text{Na-Cl} = \frac{a}{2} \approx 0.5 a$$

Equation:

$$\text{Na-Na} = \text{Cl-Cl} = \frac{a}{\sqrt{2}} \approx 0.707 a$$



Unit cell structure of a rock salt lattice. Sodium ions are shown in purple (small spheres) and chloride ions are shown in red (large spheres).

Cinnabar, named after the archetype mercury sulfide, HgS , is a distorted rock salt structure in which the resulting cell is rhombohedral (trigonal) with each atom having a coordination number of six.

Wurtzite

This is a hexagonal form of the zinc sulfide. It is identical in the number of and types of atoms, but it is built from two interpenetrating *hcp* lattices as opposed to the *fcc* lattices in zinc blende. As with zinc blende all the atoms in a wurtzite structure are 4-coordinate. The wurtzite unit cell is shown in [\[link\]](#). A number of inter atomic distances may be calculated for any material with a wurtzite cell using the lattice parameter (a).

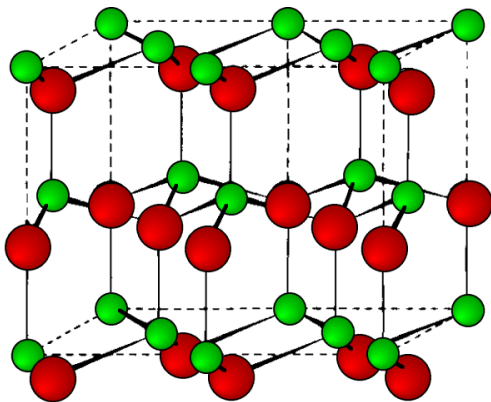
Equation:

$$\text{Zn-S} = a \sqrt{3/8} = 0.612 a = \frac{3c}{8} = 0.375 c$$

Equation:

$$\text{Zn-Zn} = \text{S-S} = a = 1.632 c$$

However, it should be noted that these formulae do not necessarily apply when the ratio a/c is different from the ideal value of 1.632.



Unit cell structure of a wurtzite lattice. Zinc atoms are shown in green (small spheres), sulfur atoms shown in red (large spheres), and the dashed lines show the unit cell.

Cesium Chloride

The cesium chloride structure is found in materials with large cations and relatively small anions. It has a simple (primitive) cubic cell ([\[link\]](#)) with a chloride ion at the corners of the cube and the cesium ion at the body center. The coordination numbers of both Cs^+ and Cl^- , with the inner atomic distances determined from the cell lattice constant (a).

Equation:

$$\text{Cs-Cl} = a \frac{\sqrt{3}}{2} \approx 0.866 a$$

Equation:

$$\text{Cs-Cs} = \text{Cl-Cl} = a$$

β -Tin.

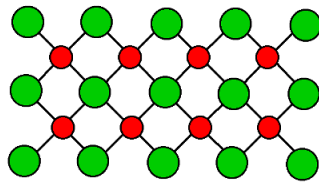
The room temperature allotrope of tin is β -tin or white tin. It has a tetragonal structure, in which each tin atom has four nearest neighbors ($\text{Sn-Sn} = 3.016 \text{ \AA}$) arranged in a very flattened tetrahedron, and two next nearest neighbors ($\text{Sn-Sn} = 3.175 \text{ \AA}$). The overall structure of β -tin consists of fused hexagons, each being linked to its neighbor via a four-membered Sn_4 ring.

Defects in crystalline solids

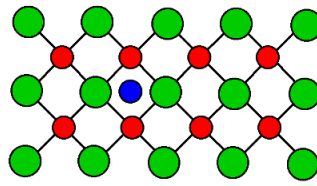
Up to this point we have only been concerned with ideal structures for crystalline solids in which each atom occupies a designated point in the crystal lattice. Unfortunately, defects ordinarily exist in equilibrium between the crystal lattice and its environment. These defects are of two general types: point defects and extended defects. As their names imply, point defects are associated with a single crystal lattice site, while extended defects occur over a greater range.

Point defects: “too many or too few” or “just plain wrong”

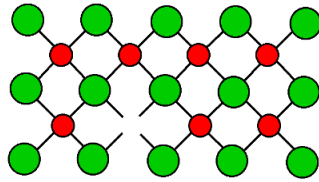
Point defects have a significant effect on the properties of a semiconductor, so it is important to understand the classes of point defects and the characteristics of each type. [\[link\]](#) summarizes various classes of native point defects, however, they may be divided into two general classes; defects with the wrong number of atoms (deficiency or surplus) and defects where the identity of the atoms is incorrect.



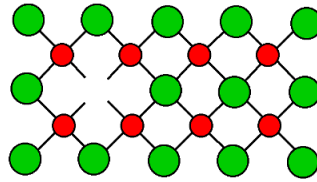
(a) perfect lattice



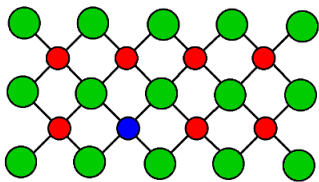
(b) interstitial impurity



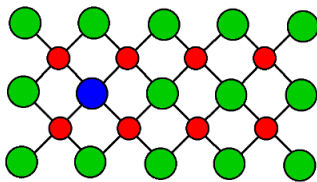
(c) cation vacancy



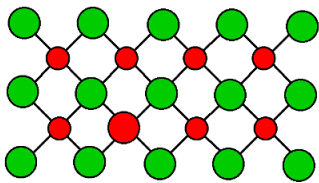
(d) anion vacancy



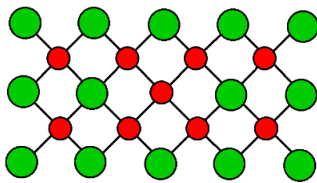
(e) substitution of cation



(f) substitution of anion



(g) B_A antisite defect



(h) A_B antisite defect

Point defects in a crystal lattice.

Interstitial Impurity

An interstitial impurity occurs when an extra atom is positioned in a lattice site that should be vacant in an ideal structure ([link](#)b). Since all the adjacent lattice sites are filled the additional atom will have to squeeze itself into the interstitial site, resulting in distortion of the lattice and alteration in the local electronic behavior of the structure. Small atoms, such as carbon,

will prefer to occupy these interstitial sites. Interstitial impurities readily diffuse through the lattice via interstitial diffusion, which can result in a change of the properties of a material as a function of time. Oxygen impurities in silicon generally are located as interstitials.

Vacancies

The converse of an interstitial impurity is when there are not enough atoms in a particular area of the lattice. These are called vacancies. Vacancies exist in any material above absolute zero and increase in concentration with temperature. In the case of compound semiconductors, vacancies can be either cation vacancies ([\[link\]](#)c) or anion vacancies ([\[link\]](#)d), depending on what type of atom are “missing”.

Substitution

Substitution of various atoms into the normal lattice structure is common, and used to change the electronic properties of both compound and elemental semiconductors. Any impurity element that is incorporated during crystal growth can occupy a lattice site. Depending on the impurity, substitution defects can greatly distort the lattice and/or alter the electronic structure. In general, cations will try to occupy cation lattice sites ([\[link\]](#)e), and anion will occupy the anion site ([\[link\]](#)f). For example, a zinc impurity in GaAs will occupy a gallium site, if possible, while a sulfur, selenium and tellurium atoms would all try to substitute for an arsenic. Some impurities will occupy either site indiscriminately, e.g., Si and Sn occupy both Ga and As sites in GaAs.

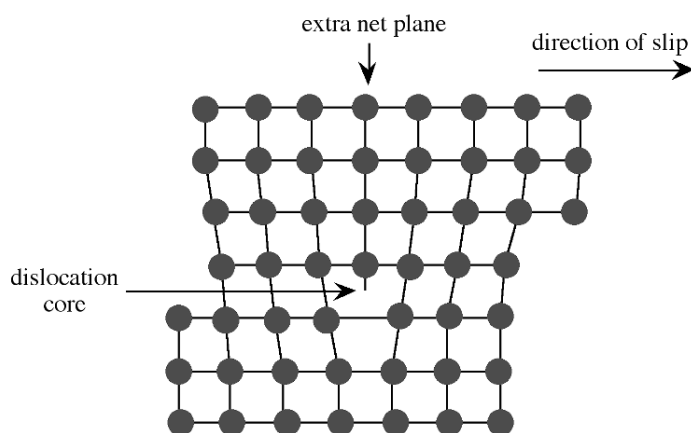
Antisite Defects

Antisite defects are a particular form of substitution defect, and are unique to compound semiconductors. An antisite defect occurs when a cation is misplaced on an anion lattice site or vice versa ([\[link\]](#)g and h). Dependant

on the arrangement these are designated as either A_B antisite defects or B_A antisite defects. For example, if an arsenic atom is on a gallium lattice site the defect would be an As_{Ga} defect. Antisite defects involve fitting into a lattice site atoms of a different size than the rest of the lattice, and therefore this often results in a localized distortion of the lattice. In addition, cations and anions will have a different number of electrons in their valence shells, so this substitution will alter the local electron concentration and the electronic properties of this area of the semiconductor.

Extended Defects: Dislocations in a Crystal Lattice

Extended defects may be created either during crystal growth or as a consequence of stress in the crystal lattice. The plastic deformation of crystalline solids does not occur such that all bonds along a plane are broken and reformed simultaneously. Instead, the deformation occurs through a dislocation in the crystal lattice. [\[link\]](#) shows a schematic representation of a dislocation in a crystal lattice. Two features of this type of dislocation are the presence of an extra crystal plane, and a large void at the dislocation core. Impurities tend to segregate to the dislocation core in order to relieve strain from their presence.



Dislocation in a crystal lattice.

Epitaxy

Epitaxy, is a transliteration of two Greek words *epi*, meaning "upon", and *taxis*, meaning "ordered". With respect to crystal growth it applies to the process of growing thin crystalline layers on a crystal substrate. In epitaxial growth, there is a precise crystal orientation of the film in relation to the substrate. The growth of epitaxial films can be done by a number of methods including molecular beam epitaxy, atomic layer epitaxy, and chemical vapor deposition, all of which will be described later.

Epitaxy of the same material, such as a gallium arsenide film on a gallium arsenide substrate, is called homoepitaxy, while epitaxy where the film and substrate material are different is called heteroepitaxy. Clearly, in homoepitaxy, the substrate and film will have the identical structure, however, in heteroepitaxy, it is important to employ where possible a substrate with the same structure and similar lattice parameters. For example, zinc selenide (zinc blende, $a = 5.668 \text{ \AA}$) is readily grown on gallium arsenide (zinc blende, $a = 5.653 \text{ \AA}$). Alternatively, epitaxial crystal growth can occur where there exists a simple relationship between the structures of the substrate and crystal layer, such as is observed between Al_2O_3 (100) on Si (100). Whichever route is chosen a close match in the lattice parameters is required, otherwise, the strains induced by the lattice mismatch results in distortion of the film and formation of dislocations. If the mismatch is significant epitaxial growth is not energetically favorable, causing a textured film or polycrystalline untextured film to be grown. As a general rule of thumb, epitaxy can be achieved if the lattice parameters of the two materials are within about 5% of each other. For good quality epitaxy, this should be less than 1%. The larger the mismatch, the larger the strain in the film. As the film gets thicker and thicker, it will try to relieve the strain in the film, which could include the loss of epitaxy of the growth of dislocations. It is important to note that the $\langle 100 \rangle$ directions of a film must be parallel to the $\langle 100 \rangle$ direction of the substrate. In some cases, such as Fe on MgO, the [111] direction is parallel to the substrate [100]. The epitaxial relationship is specified by giving first the plane in the film that is parallel to the substrate [100].

Bibliography

- *International Tables for X-ray Crystallography*. Vol. IV; Kynoch Press: Birmingham, UK (1974).
- B. F. G. Johnson, in *Comprehensive Inorganic Chemistry*, Pergamon Press, Vol. 4, Chapter 52 (1973).
- A. R. West, *Solid State Chemistry and its Applications*, Wiley, New York (1984).

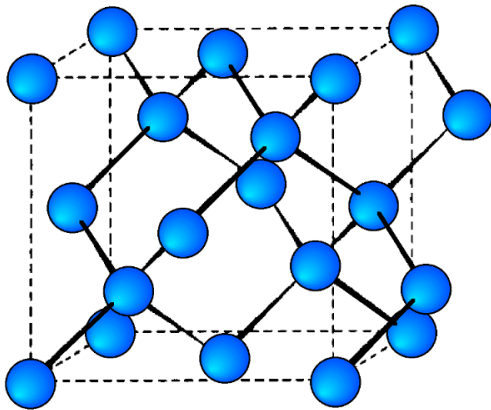
Structures of Element and Compound Semiconductors

Introduction

A single crystal of either an elemental (e.g., silicon) or compound (e.g., gallium arsenide) semiconductor forms the basis of almost all semiconductor devices. The ability to control the electronic and optoelectronic properties of these materials is based on an understanding of their structure. In addition, the metals and many of the insulators employed within a microelectronic device are also crystalline.

Group IV (14) elements

Each of the semiconducting phases of the group IV (14) elements, C (diamond), Si, Ge, and α -Sn, adopt the diamond cubic structure ([\[link\]](#)). Their lattice constants (a , Å) and densities (ρ , g/cm³) are given in [\[link\]](#).



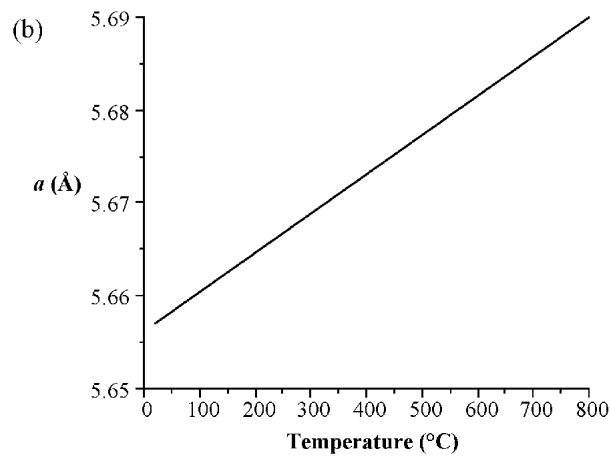
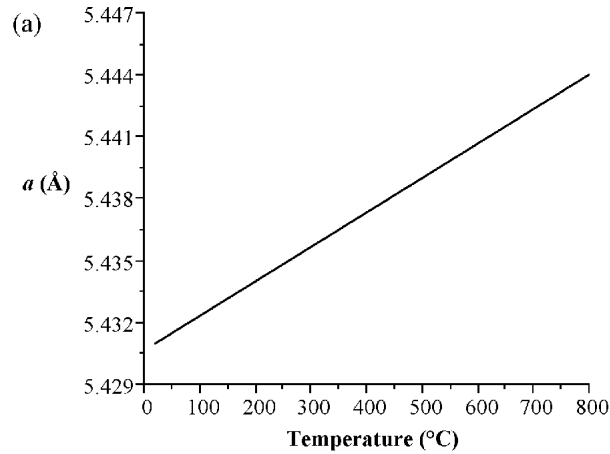
Unit cell structure of a diamond cubic lattice showing the two interpenetrating face-centered cubic lattices.

Element	Lattice parameter, a (Å)	Density (g/cm ³)
carbon (diamond)	3.56683(1)	3.51525
silicon	5.4310201(3)	2.319002
germanium	5.657906(1)	5.3234
tin (α -Sn)	6.4892(1)	7.285

Lattice parameters and densities (measured at 298 K) for the diamond cubic forms of the group IV (14) elements.

As would be expected the lattice parameter increase in the order C < Si < Ge < α -Sn. Silicon and germanium form a continuous series of solid solutions with gradually varying parameters. It is worth noting the high degree of accuracy that the lattice parameters are known for high purity crystals of these elements. In addition, it is important to note the temperature at which structural measurements are made, since the lattice parameters are temperature dependent ([\[link\]](#)). The lattice constant (a), in Å, for high purity silicon may be calculated for any temperature (T) over the temperature range 293 - 1073 K by the formula shown below.

$$a_T = 5.4304 + 1.8138 \times 10^{-5} (T - 298.15 \text{ K}) + 1.542 \times 10^{-9} (T - 298.15 \text{ K})^2$$



Temperature dependence of the
lattice parameter for (a) Si and
(b) Ge.

Even though the diamond cubic forms of Si and Ge are the only forms of direct interest to semiconductor devices, each exists in numerous crystalline high pressure and meta-stable forms. These are described along with their interconversions, in [\[link\]](#).



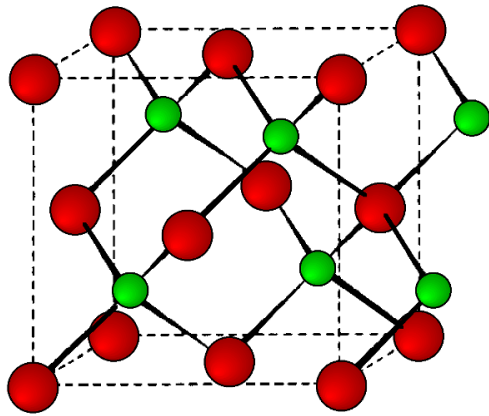
Phase	Structure	Remarks
Si I	diamond cubic	stable at normal pressure
Si II	grey tin structure	formed from Si I or Si V above 14 GPa
Si III	cubic	metastable, formed from Si II above 10 GPa
Si IV	hexagonal	
Si V	unidentified	stable above 34 GPa, formed from Si II above 16 GPa
Si VI	hexagonal close packed	stable above 45 GPa
Ge I	diamond cubic	low-pressure phase
Ge II	β -tin structure	formed from Ge I above 10 GPa
Ge III	tetragonal	formed by quenching Ge II at low pressure
Ge IV	body centered cubic	formed by quenching Ge II to 1 atm at 200 K

High pressure and metastable phases of silicon and germanium.

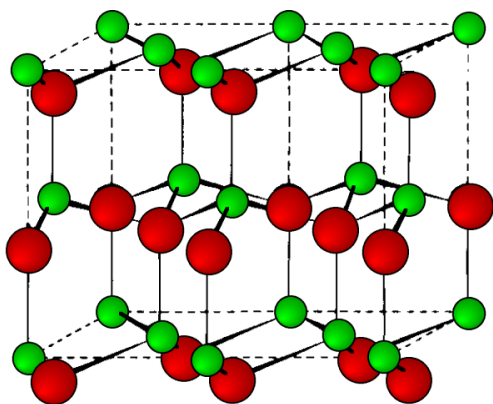
Group III-V (13-15) compounds

The stable phases for the arsenides, phosphides and antimonides of aluminum, gallium and indium all exhibit zinc blende structures ([\[link\]](#)). In contrast, the nitrides are found as wurtzite structures (e.g., [\[link\]](#)). The structure, lattice parameters, and densities of the III-V compounds are given

in [\[link\]](#). It is worth noting that contrary to expectation the lattice parameter of the gallium compounds is smaller than their aluminum homolog; for GaAs $a = 5.653 \text{ \AA}$; AlAs $a = 5.660 \text{ \AA}$. As with the group IV elements the lattice parameters are highly temperature dependent; however, additional variation arises from any deviation from absolute stoichiometry. These effects are shown in [\[link\]](#).



Unit cell structure of a zinc blende (ZnS) lattice. Zinc atoms are shown in green (small), sulfur atoms shown in red (large), and the dashed lines show the unit cell.

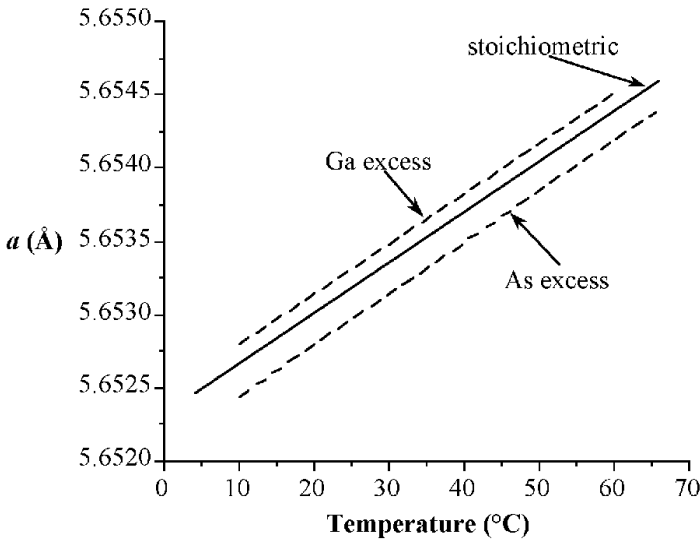


Unit cell structure of a wurtzite lattice. Zinc atoms are shown in green (small), sulfur atoms shown in red (large), and the dashed lines show the unit cell.

Compound	Structure	Lattice parameter (Å)	Density (g/cm ³)
AlN	wurtzite	$a = 3.11(1)$, $c = 4.98(1)$	3.255
AlP	zinc blende	$a = 5.4635(4)$	2.40(1)
AlAs	zinc blende	$a = 5.660$	3.760

AlSb	zinc blende	$a = 6.1355(1)$	4.26
GaN	wurtzite	$a = 3.190, c = 5.187$	
GaP	zinc blende	$a = 5.4505(2)$	4.138
GaAs	zinc blende	$a = 5.65325(2)$	5.3176(3)
InN	wurtzite	$a = 3.5446, c = 5.7034$	6.81
InP	zinc blende	$a = 5.868(1)$	4.81
InAs	zinc blende	$a = 6.0583$	5.667
InSb	zinc blende	$a = 6.47937$	5.7747(4)

Lattice parameters and densities (measured at 298 K) for the III-V (13-15) compound semiconductors. Estimated standard deviations given in parentheses.



Temperature dependence of the lattice parameter for stoichiometric GaAs and crystals with either Ga or As excess.

The homogeneity of structures of alloys for a wide range of solid solutions to be formed between III-V compounds in almost any combination. Two classes of ternary alloys are formed: $\text{III}_x\text{-III}_{1-x}\text{-V}$ (e.g., $\text{Al}_x\text{-Ga}_{1-x}\text{-As}$) and $\text{III-V}_{1-x}\text{-V}_x$ (e.g., $\text{Ga-As}_{1-x}\text{-P}_x$). While quaternary alloys of the type $\text{III}_x\text{-III}_{1-x}\text{-V}_y\text{-V}_{1-y}$ allow for the growth of materials with similar lattice parameters, but a broad range of band gaps. A very important ternary alloy, especially in optoelectronic applications, is $\text{Al}_x\text{-Ga}_{1-x}\text{-As}$ and its lattice parameter (a) is directly related to the composition (x).

$$a = 5.6533 + 0.0078 x$$

Not all of the III-V compounds have well characterized high-pressure phases. However, in each case where a high-pressure phase is observed the coordination number of both the group III and group V element increases from four to six. Thus, AlP undergoes a zinc blende to rock salt transformation at high pressure above 170 kbar, while AlSb and GaAs form orthorhombic distorted rock salt structures above 77 and 172 kbar,

respectively. An orthorhombic structure is proposed for the high-pressure form of InP (>133 kbar). Indium arsenide (InAs) undergoes two-phase transformations. The zinc blende structure is converted to a rock salt structure above 77 kbar, which in turn forms a β -tin structure above 170 kbar.

Group II-VI (12-16) compounds

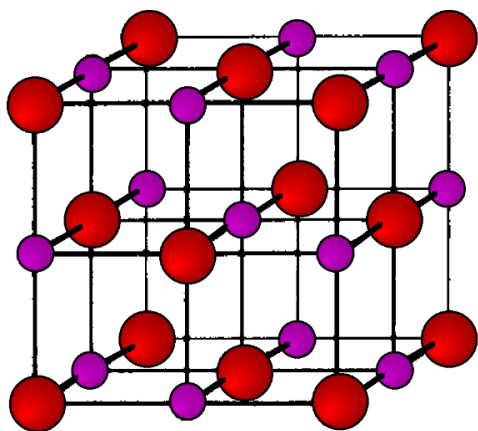
The structures of the II-VI compound semiconductors are less predictable than those of the III-V compounds (above), and while zinc blende structure exists for almost all of the compounds there is a stronger tendency towards the hexagonal wurtzite form. In several cases the zinc blende structure is observed under ambient conditions, but may be converted to the wurtzite form upon heating. In general the wurtzite form predominates with the smaller anions (e.g., oxides), while the zinc blende becomes the more stable phase for the larger anions (e.g., tellurides). One exception is mercury sulfide (HgS) that is the archetype for the trigonal cinnabar phase. [\[link\]](#) lists the stable phase of the chalcogenides of zinc, cadmium and mercury, along with their high temperature phases where applicable. Solid solutions of the II-VI compounds are not as easily formed as for the III-V compounds; however, two important examples are $\text{ZnS}_x\text{Se}_{1-x}$ and $\text{Cd}_x\text{Hg}_{1-x}\text{Te}$.

Compound	Structure	Lattice parameter (Å)	Density (g/cm ³)
ZnS	zinc blende	$a = 5.410$	4.075
	wurtzite	$a = 3.822, c = 6.260$	4.087

ZnSe	Zinc blende	$a = 5.668$	5.27
ZnTe	Zinc blende	$a = 6.10$	5.636
CdS	wurtzite	$a = 4.136, c = 6.714$	4.82
CdSe	wurtzite	$a = 4.300, c = 7.011$	5.81
CdTe	Zinc blende	$a = 6.482$	5.87
HgS	cinnabar	$a = 4.149, c = 9.495$	
	Zinc blende	$a = 5.851$	7.73
HgSe	Zinc blende	$a = 6.085$	8.25
HgTe	Zinc blende	$a = 6.46$	8.07

Lattice parameters and densities (measured at 298 K) for the II-VI (12-16) compound semiconductors.

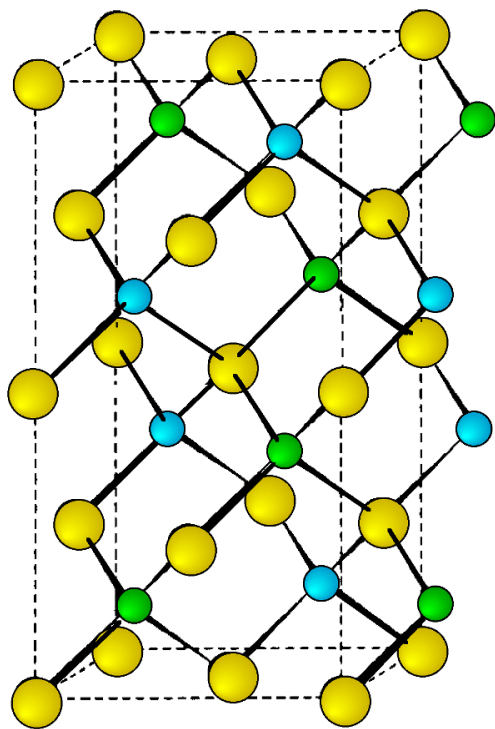
The zinc chalcogenides all transform to a cesium chloride structure under high pressures, while the cadmium compounds all form rock salt high-pressure phases ([\[link\]](#)). Mercury selenide (HgSe) and mercury telluride (HgTe) convert to the mercury sulfide archetype structure, cinnabar, at high pressure.



Unit cell structure of a rock salt lattice. Sodium ions are shown in purple and chloride ions are shown in red.

I-III-VI₂ (11-13-16) compounds

Nearly all I-III-VI₂ compounds at room temperature adopt the chalcopyrite structure ([\[link\]](#)). The cell constants and densities are given in [\[link\]](#). Although there are few reports of high temperature or high-pressure phases, AgInS₂ has been shown to exist as a high temperature orthorhombic polymorph ($a = 6.954$, $b = 8.264$, and $c = 6.683$ Å), and AgInTe₂ forms a cubic phase at high pressures.



Unit cell structure of a chalcopyrite lattice. Copper atoms are shown in blue, iron atoms are shown in green and sulfur atoms are shown in yellow. The dashed lines show the unit cell.

Compound	Lattice parameter a (Å)	Lattice parameter c (Å)	Density (g.cm ³)

CuAlS ₂	5.32	10.430	3.45
CuAlSe ₂	5.61	10.92	4.69
CuAlTe ₂	5.96	11.77	5.47
CuGaS ₂	5.35	10.46	4.38
CuGaSe ₂	5.61	11.00	5.57
CuGaTe ₂	6.00	11.93	5.95
CuInS ₂	5.52	11.08	4.74
CuInSe ₂	5.78	11.55	5.77
CuInTe ₂	6.17	12.34	6.10
AgAlS ₂	6.30	11.84	6.15
AgGaS ₂	5.75	10.29	4.70
AgGaSe ₂	5.98	10.88	5.70
AgGaTe ₂	6.29	11.95	6.08
AgInS ₂	5.82	11.17	4.97
AgInSe ₂	6.095	11.69	5.82
AgInTe ₂	6.43	12.59	6.96

Chalcopyrite lattice parameters and densities (measured at 298 K) for the I-III-VI compound semiconductor. Lattice parameters for tetragonal cell.

Of the I-III-VI₂ compounds, the copper indium chalcogenides (CuInE₂) are certainly the most studied for their application in solar cells. One of the

advantages of the copper indium chalcogenide compounds is the formation of solid solutions (alloys) of the formula $\text{CuInE}_{2-x}\text{E}'_x$, where the composition variable (x) varies from 0 to 2. The $\text{CuInS}_{2-x}\text{Se}_x$ and $\text{CuInSe}_{2-x}\text{Te}_x$ systems have also been examined, as has the $\text{CuGa}_y\text{In}_{1-y}\text{S}_{2-x}\text{Se}_x$ quaternary system. As would be expected from a consideration of the relative ionic radii of the chalcogenides the lattice parameters of the $\text{CuInS}_{2-x}\text{Se}_x$ alloy should increase with increased selenium content. Vergard's law requires the lattice constant for a linear solution of two semiconductors to vary linearly with composition (e.g., as is observed for $\text{Al}_x\text{Ga}_{1-x}\text{As}$), however, the variation of the tetragonal lattice constants (a and c) with composition for $\text{CuInS}_{2-x}\text{S}_x$ are best described by the parabolic relationships.

$$a = 5.532 + 0.0801 x + 0.0260 x^2$$

$$c = 11.156 + 0.1204 x + 0.0611 x^2$$

A similar relationship is observed for the $\text{CuInSe}_{2-x}\text{Te}_x$ alloys.

$$a = 5.783 + 0.1560 x + 0.0212 x^2$$

$$c = 11.628 + 0.3340 x + 0.0277 x^2$$

The large difference in ionic radii between S and Te (0.37 Å) prevents formation of solid solutions in the $\text{CuInS}_{2-x}\text{Te}_x$ system, however, the single alloy $\text{CuInS}_{1.5}\text{Te}_{0.5}$ has been reported.

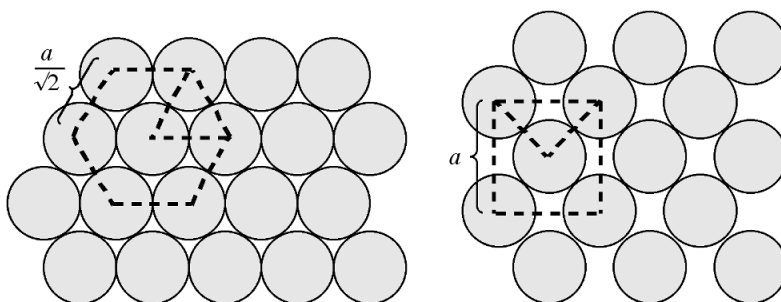
Orientation effects

Once single crystals of high purity silicon or gallium arsenide are produced they are cut into wafers such that the exposed face of these wafers is either the crystallographic {100} or {111} planes. The relative structure of these surfaces are important with respect to oxidation, etching and thin film growth. These processes are orientation-sensitive; that is, they depend on the direction in which the crystal slice is cut.

Atom density and dangling bonds

The principle planes in a crystal may be differentiated in a number of ways, however, the atom and/or bond density are useful in predicting much of the chemistry of semiconductor surfaces. Since both silicon and gallium arsenide are *fcc* structures and the {100} and {111} are the only technologically relevant surfaces, discussions will be limited to *fcc* {100} and {111}.

The atom density of a surface may be defined as the number of atoms per unit area. [\[link\]](#) shows a schematic view of the {111} and {100} planes in a *fcc* lattice. The {111} plane consists of a hexagonal close packed array in which the crystal directions within the plane are oriented at 60° to each other. The hexagonal packing and the orientation of the crystal directions are indicated in [\[link\]](#)b as an overlaid hexagon. Given the intra-planar inter-atomic distance may be defined as a function of the lattice parameter, the area of this hexagon may be readily calculated. For example in the case of silicon, the hexagon has an area of 38.30 Å². The number of atoms within the hexagon is three: the atom in the center plus 1/3 of each of the six atoms at the vertices of the hexagon (each of the atoms at the hexagons vertices is shared by three other adjacent hexagons). Thus, the atom density of the {111} plane is calculated to be 0.0783 Å⁻². Similarly, the atom density of the {100} plane may be calculated. The {100} plane consists of a square array in which the crystal directions within the plane are oriented at 90° to each other. Since the square is coincident with one of the faces of the unit cell the area of the square may be readily calculated. For example in the case of silicon, the square has an area of 29.49 Å². The number of atoms within the square is 2: the atom in the center plus 1/4 of each of the four atoms at the vertices of the square (each of the atoms at the corners of the square are shared by four other adjacent squares). Thus, the atom density of the {100} plane is calculated to be 0.0678 Å⁻². While these values for the atom density are specific for silicon, their ratio is constant for all diamond cubic and zinc blende structures: {100}:{111} = 1:1.155. In general, the fewer dangling bonds the more stable a surface structure.

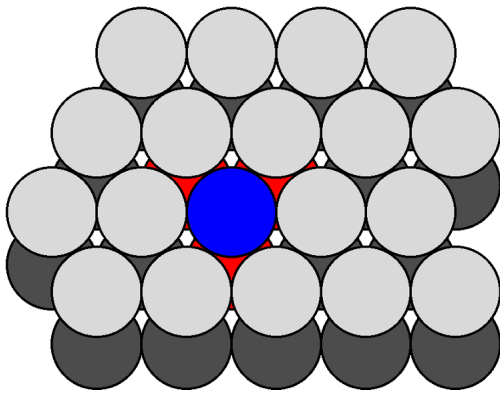


Schematic representation of the (111) and (100) faces of a face centered cubic (fcc) lattice showing the relationship between the close packed rows.

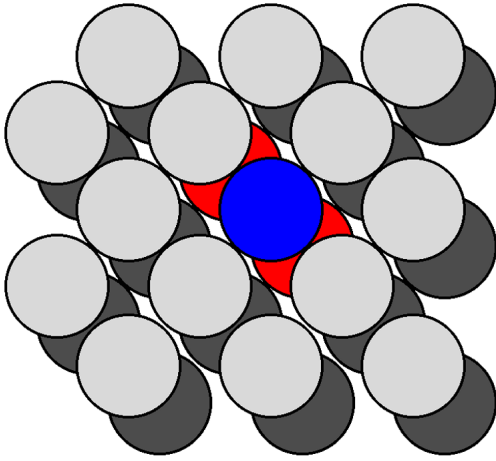
An atom inside a crystal of any material will have a coordination number (n) determined by the structure of the material. For example, all atoms within the bulk of a silicon crystal will be in a tetrahedral four-coordinate environment ($n = 4$). However, at the surface of a crystal the atoms will not make their full complement of bonds. Each atom will therefore have less nearest neighbors than an atom within the bulk of the material. The missing bonds are commonly called dangling bonds. While this description is not particularly accurate it is, however, widely employed and as such will be used herein. The number of dangling bonds may be defined as the difference between the ideal coordination number (determined by the bulk crystal structure) and the actual coordination number as observed at the surface.

[\[link\]](#) shows a section of the $\{111\}$ surfaces of a diamond cubic lattice viewed perpendicular to the $\{111\}$ plane. The atoms within the bulk have a coordination number of four. In contrast, the atoms at the surface (e.g., the atom shown in blue in [\[link\]](#)) are each bonded to just three other atoms (the atoms shown in red in [\[link\]](#)), thus each surface atom has one dangling bond. As can be seen from [\[link\]](#), which shows the atoms at the $\{100\}$ surface viewed perpendicular to the $\{100\}$ plane, each atom at the surface (e.g., the atom shown in blue in [\[link\]](#)) is only coordinated to two other atoms (the atoms shown in red in [\[link\]](#)), leaving two dangling bonds per

atom. It should be noted that the same number of dangling bonds are found for the $\{111\}$ and $\{100\}$ planes of a zinc blende lattice. The ratio of dangling bonds for the $\{100\}$ and $\{111\}$ planes of all diamond cubic and zinc blende structures is $\{100\}:\{111\} = 2:1$. Furthermore, since the atom densities of each plane are known then the ratio of the dangling bond densities is determined to be: $\{100\}:\{111\} = 1:0.577$.



A section of the $\{111\}$ surfaces of a diamond cubic lattice viewed perpendicular to the $\{111\}$ plane.



A section of the $\{100\}$ surface of a diamond cubic lattice viewed perpendicular to the $\{100\}$ plane.

Silicon

For silicon, the $\{111\}$ planes are closer packed than the $\{100\}$ planes. As a result, growth of a silicon crystal is therefore slowest in the $\langle 111 \rangle$ direction, since it requires laying down a close packed atomic layer upon another layer in its closest packed form. As a consequence $\langle 111 \rangle$ Si is the easiest to grow, and therefore the least expensive.

The dissolution or etching of a crystal is related to the number of broken bonds already present at the surface: the fewer bonds to be broken in order to remove an individual atom from a crystal, the easier it will be to dissolve the crystal. As a consequence of having only one dangling bond (requiring three bonds to be broken) etching silicon is slowest in the $\langle 111 \rangle$ direction. The electronic properties of a silicon wafer are also related to the number of dangling bonds.

Silicon microcircuits are generally formed on a single crystal wafer that is diced after fabrication by either sawing part way through the wafer thickness or scoring (scribing) the surface, and then physically breaking. The physical breakage of the wafer occurs along the natural cleavage planes, which in the case of silicon are the $\{111\}$ planes.

Gallium arsenide

The zinc blende lattice observed for gallium arsenide results in additional considerations over that of silicon. Although the $\{100\}$ plane of GaAs is structurally similar to that of silicon, two possibilities exist: a face consisting of either all gallium atoms or all arsenic atoms. In either case the surface atoms have two dangling bonds, and the properties of the face are independent of whether the face is gallium or arsenic.

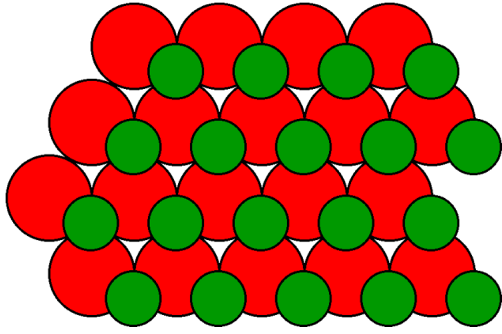
The $\{111\}$ plane also has the possibility of consisting of all gallium or all arsenic. However, unlike the $\{100\}$ planes there is a significant difference between the two possibilities. [\[link\]](#) shows the gallium arsenide structure represented by two interpenetrating *fcc* lattices. The $[111]$ axis is vertical within the plane of the page. Although the structure consists of alternate layers of gallium and arsenic stacked along the $[111]$ axis, the distance between the successive layers alternates between large and small. Assigning arsenic as the parent lattice the order of the layers in the $[111]$ direction is

As — Ga — As — Ga — As — Ga, while in the $\left[\begin{smallmatrix} \bar{\bar{\bar{1}}} & \bar{\bar{\bar{1}}} & \bar{\bar{\bar{1}}} \end{smallmatrix}\right]$ direction the layers are

ordered, Ga — As — Ga — As — Ga — As ([\[link\]](#)). In silicon these two directions are of course identical. The surface of a crystal would be either arsenic, with three dangling bonds, or gallium, with one dangling bond. Clearly, the latter is energetically more favorable. Thus, the (111) plane shown in [\[link\]](#) is

called the (111) Ga face. Conversely, the $\left[\begin{smallmatrix} \bar{\bar{\bar{1}}} & \bar{\bar{\bar{1}}} & \bar{\bar{\bar{1}}} \end{smallmatrix}\right]$ plane would be either gallium, with three dangling bonds, or arsenic, with one dangling bond.

Again, the latter is energetically more favorable and the $\left[\begin{smallmatrix} \bar{\bar{\bar{1}}} & \bar{\bar{\bar{1}}} & \bar{\bar{\bar{1}}} \end{smallmatrix}\right]$ plane is therefore called the (111) As face.



The (111) Ga face of GaAs showing a surface layer containing gallium atoms (green) with one dangling bond per gallium and three bonds to the arsenic atoms (red) in the lower layer.

The (111) As is distinct from that of (111) Ga due to the difference in the number of electrons at the surface. As a consequence, the (111) As face etches more rapidly than the (111) Ga face. In addition, surface evaporation below 770 °C occurs more rapidly at the (111) As face.

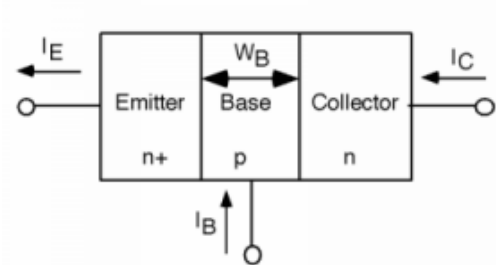
Bibliography

- M. Baublitz and A. L. Ruoff, *J. Appl. Phys.*, 1982, **53**, 6179.
- J. C. Jamieson, *Science*, 1963, **139**, 845.
- C. C. Landry, J. Lockwood, and A. R. Barron, *Chem. Mater.*, 1995, **7**, 699.
- M. Robbins, J. C. Phillips, and V. G. Lambrecht, *J. Phys. Chem. Solids*, 1973, **34**, 1205.
- D. Sridevi and K. V. Reddy, *Mat. Res. Bull.*, 1985, **20**, 929.
- Y. K. Vohra, S. T. Weir, and A. L. Ruoff, *Phys. Rev. B*, 1985, **31**, 7344.
- W. M. Yin and R. J. Paff, *J. Appl. Phys.*, 1973, **45**, 1456.

Introduction to Bipolar Transistors

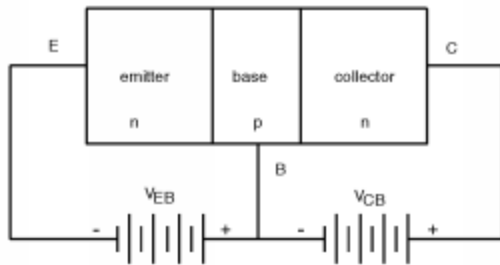
Note: This module is adapted from the Connexions module entitled *Introduction to Bipolar Transistors* by Bill Wilson.

Let's leave the world of two terminal devices (which are all called diodes by the way; diode just means two-terminals) and venture into the much more interesting world of three terminals. The first device we will look at is called the *bipolar transistor*. Consider the structure shown in [\[link\]](#):



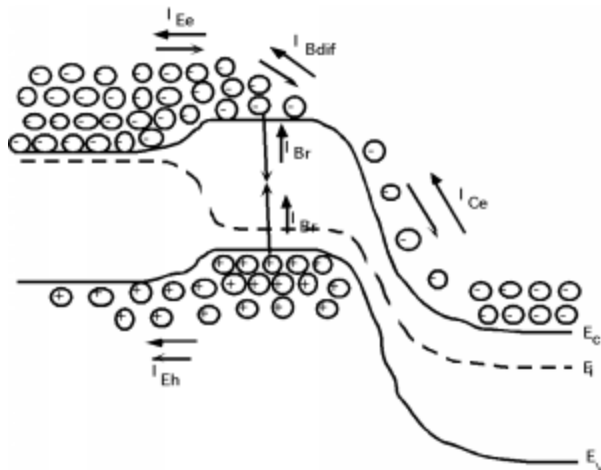
Structure of a npn bipolar transistor.

The device consists of three layers of silicon, a heavily doped n-type layer called the emitter, a moderately doped p-type layer called the base, and third, more lightly doped layer called the collector. In a biasing (applied DC potential) configuration called *forward active biasing*, the emitter-base junction is forward biased, and the base-collector junction is reverse biased. [\[link\]](#) shows the biasing conventions we will use. Both bias voltages are referenced to the base terminal. Since the base-emitter junction is forward biased, and since the base is made of p-type material, V_{EB} must be negative. On the other hand, in order to reverse bias the base-collector junction V_{CB} will be a positive voltage.



Forward active biasing of a npn bipolar transistor.

Now, let's draw the band-diagram for this device. At first this might seem hard to do, but we know what forward and reverse biased band diagrams look like, so we'll just stick one of each together. We show this in [\[link\]](#), which is a very busy figure, but it is also very important, because it shows all of the important features in the operation the transistor. Since the base-emitter junction is forward biased, electrons will go from the (n-type) emitter into the base. Likewise, some holes from the base will be injected into the emitter.



Band diagram and carrier fluxes in a bipolar transistor.

In [\[link\]](#), we have two different kinds of arrows. The open arrows which are attached to the carriers, show us which way the carrier is moving. The solid arrows which are labeled with some kind of subscripted I , represent current flow. We need to do this because for holes, motion and current flow are in the same direction, while for electrons, carrier motion and current flow are in opposite directions.

Just as we saw in the last chapter, the electrons which are injected into the base diffuse away from the emitter-base junction towards the (reverse biased) base-collector junction. As they move through the base, some of the electrons encounter holes and recombine with them. Those electrons which do get to the base-collector junction run into a large electric field which sweeps them out of the base and into the collector. They "fall" down the large potential drop at the junction.

These effects are all seen in [\[link\]](#), with arrows representing the various currents which are associated with each of the carriers fluxes. I_{Ee} represents the current associated with the electron injection into the base, i.e., it points in the opposite direction from the motion of the electrons, since electrons have a negative charge. I_{Eh} represents the current associated with holes injection into the emitter from the base. I_{Br} represents recombination current in the base, while I_{Ce} represents the electron current going into the collector. It should be easy for you to see that:

Equation:

$$I_E = I_{Ee} + I_{Eh}$$

Equation:

$$I_B = I_{Eh} + I_{Br}$$

Equation:

$$I_C = I_{Ce}$$

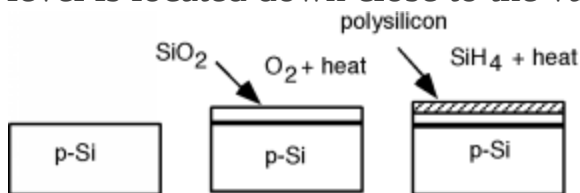
In a "good" transistor, almost all of the current across the base-emitter junction consists of electrons being injected into the base. The transistor engineer works hard to design the device so that very little emitter current is

made up of holes coming from the base into the emitter. The transistor is also designed so that almost all of those electrons which are injected into the base make it across to the base-collector reverse-biased junction. Some recombination is unavoidable, but things are arranged so as to minimize this effect.

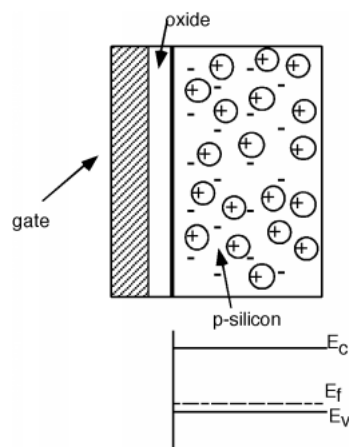
Basic MOS Structure

Note: This module is adapted from the Connexions module entitled *Basic MOS Structure* by Bill Wilson.

[\[link\]](#) shows the basic steps necessary to make the MOS structure. It will help us in our understanding if we now rotate our picture so that it is pointing sideways in our next few drawings. [\[link\]](#) shows the rotated structure. Note that in the p-silicon we have positively charged mobile holes, and negatively charged, fixed acceptors. Because we will need it later, we have also shown the band diagram for the semiconductor below the sketch of the device. Note that since the substrate is p-type, the Fermi level is located down close to the valance band.

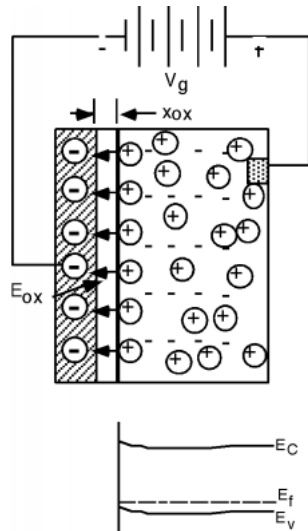


Formation of the metal-oxide-semiconductor (MOS) structure.



Basic metal-oxide-semiconductor (MOS) structure.

Let us now place a potential between the gate and the silicon substrate. Suppose we make the gate negative with respect to the substrate. Since the substrate is p-type, it has a lot of mobile, positively charged holes in it. Some of them will be attracted to the negative charge on the gate, and move over to the surface of the substrate. This is also reflected in the band diagram shown in [\[link\]](#). Remember that the density of holes is exponentially proportional to how close the Fermi level is to the valence band edge. We see that the band diagram has been bent up slightly near the surface to reflect the extra holes which have accumulated there.



Applying a negative gate voltage to a basic metal-oxide-semiconductor (MOS) structure.

An electric field will develop between the positive holes and the negative gate charge. Note that the gate and the substrate form a kind of parallel plate capacitor, with the oxide acting as the insulating layer in-between them. The oxide is quite thin compared to the area of the device, and so it is quite appropriate to assume that the electric field inside the oxide is a uniform one. (We will ignore fringing at the edges.) The integral of the electric field is just the applied gate voltage V_g . If the oxide has a thickness x_{ox} then since E_{ox} is uniform, it is given by, [\[link\]](#).

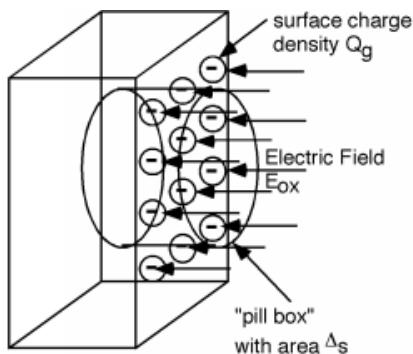
Equation:

$$E_{ox} = \frac{V_g}{x_{ox}}$$

If we focus in on a small part of the gate, we can make a little "pill" box which extends from somewhere in the oxide, across the oxide/gate interface and ends up inside the gate material someplace. The pill-box will have an area Δs . Now we will invoke Gauss' law which we reviewed earlier. Gauss' law simply says that the surface integral over a closed surface of the displacement vector D (which is, of course, $\epsilon \times E$) is equal to the total charge enclosed by that surface. We will assume that there is a surface charge density $-Q_g$ Coulombs/cm² on the surface of the gate electrode ([\[link\]](#)). The integral form of Gauss' Law is just:

Equation:

$$\oint \epsilon_{ox} \mathbf{E} \, d\mathbf{S} = Q_{encl}$$



Finding the surface
charge density.

Note that we have used $\epsilon_{\text{ox}}E$ in place of D . In this particular set-up the integral is easy to perform, since the electric field is uniform, and only pointing in through one surface - it terminates on the negative surface charge inside the pill-box. The charge enclosed in the pill box is just - ($Q_g\Delta s$), and so we have (keeping in mind that the surface integral of a vector pointing into the surface is negative), [\[link\]](#), or [\[link\]](#).

Equation:

$$\oint \epsilon_{\text{ox}} \mathbf{E} \cdot d\mathbf{S} = -(\epsilon_{\text{ox}} E_{\text{ox}} \Delta(s)) \\ = -(Q_g \Delta(s))$$

Equation:

$$\epsilon_{\text{ox}} E_{\text{ox}} = Q_g$$

Now, we can use [\[link\]](#) to get [\[link\]](#) or [\[link\]](#).

Equation:

$$\frac{\epsilon_{\text{ox}} V_g}{x_{\text{ox}}} = Q_g$$

Equation:

$$\frac{Q_g}{V_g} = \frac{\epsilon_{\text{ox}}}{x_{\text{ox}}} \equiv c_{\text{ox}}$$

The quantity c_{ox} is called the oxide capacitance. It has units of Farads/cm², so it is really a capacitance per unit area of the oxide. The dielectric constant of silicon dioxide, ϵ_{ox} , is about 3.3×10^{-13} F/cm. A typical oxide thickness might be 250 Å (or 2.5×10^{-6} cm). In this case, c_{ox} would be about 1.30×10^{-7} F/cm². The units we are using here, while they might

seem a little arbitrary and confusing, are the ones most commonly used in the semiconductor business.

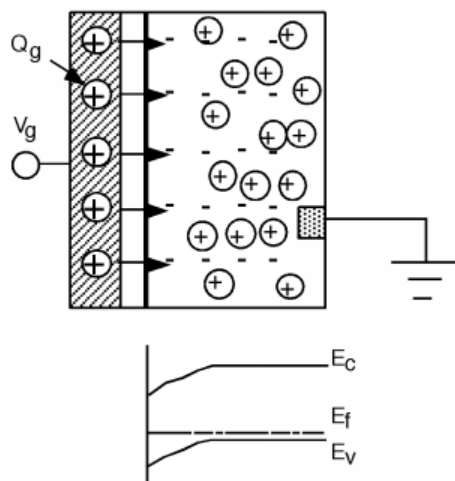
The most useful form of [\[link\]](#) is when it is turned around, [\[link\]](#), as it gives us a way to find the charge on the gate in terms of the gate potential. We will use this equation later in our development of how the MOS transistor really works.

Equation:

$$Q_g = c_{\text{ox}} V_g$$

It turns out we have not done anything very useful by apply a negative voltage to the gate. We have drawn more holes there in what is called an accumulation layer, but that is not helping us in our effort to create a layer of electrons in the MOSFET which could electrically connect the two n-regions together.

Let's turn the battery around and apply a positive voltage to the gate ([\[link\]](#)). Actually, let's take the battery out for now, and just let V_g be a positive value, relative to the substrate which will tie to ground. Making V_g positive puts positive Q_g on the gate. The positive charge pushes the holes away from the region under the gate and uncovers some of the negatively-charged fixed acceptors. Now the electric field points the other way, and goes from the positive gate charge, terminating on the negative acceptor charge within the silicon.

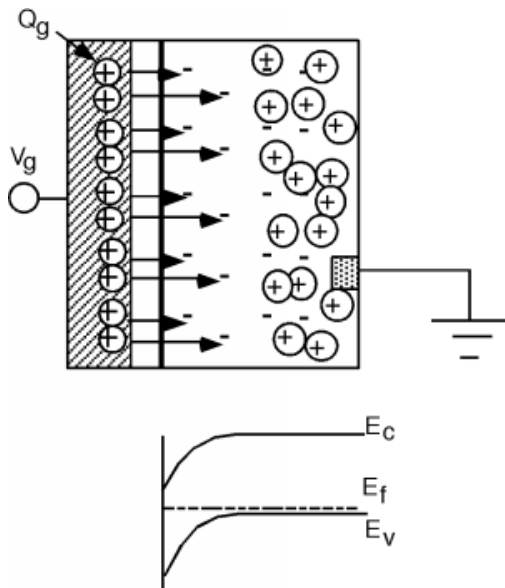


Increasing the voltage extends the depletion region further into the device.

The electric field now extends into the semiconductor. We know from our experience with the p-n junction that when there is an electric field, there is a shift in potential, which is represented in the band diagram by bending the bands. Bending the bands down (as we should moving towards positive charge) causes the valence band to pull away from the Fermi level near the surface of the semiconductor. If you remember the expression we had for the density of holes in terms of E_v and E_f it is easy to see that indeed, [\[link\]](#), there is a depletion region (region with almost no holes) near the region under the gate. (Once $E_f - E_v$ gets large with respect to kT , the negative exponent causes $p \rightarrow 0$.)

Equation:

$$p = N_v e^{-\frac{E_f - E_v}{kT}}$$



Threshold, E_f is getting close to E_c .

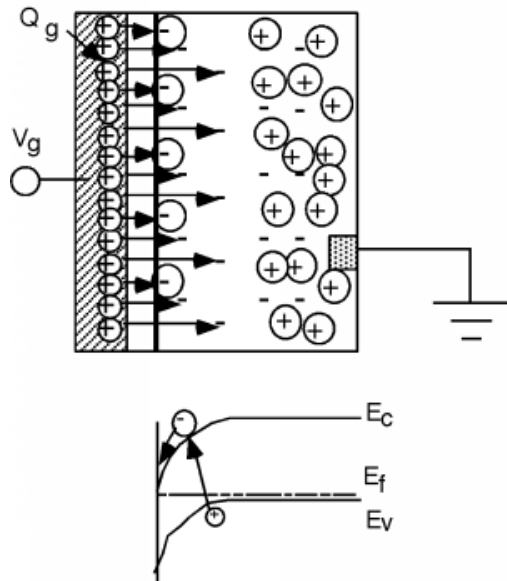
The electric field extends further into the semiconductor, as more negative charge is uncovered and the bands bend further down. But now we have to recall the electron density equation, which tells us how many electrons we have:

Equation:

$$n = N_c e^{-\frac{E_c - E_f}{kT}}$$

A glance at [\[link\]](#) reveals that with this much band bending, E_c the conduction band edge, and E_f the Fermi level are starting to get close to one another (at least compared to kT), which means that n , the electron concentration, should soon start to become significant. In the situation represented by [\[link\]](#), we say we are at *threshold*, and the gate voltage at this point is called the *threshold voltage*, V_T .

Now, let's increase V_g above V_T . Here's the sketch in [\[link\]](#). Even though we have increased V_g beyond the threshold voltage, V_T , and more positive charge appears on the gate, the depletion region no longer moves back into the substrate. Instead electrons start to appear under the gate region, and the additional electric field lines terminate on these new electrons, instead of on additional acceptors. We have created an *inversion layer* of electrons under the gate, and it is this layer of electrons which we can use to connect the two n-type regions in our initial device.



Inversion - electrons form
under the gate.

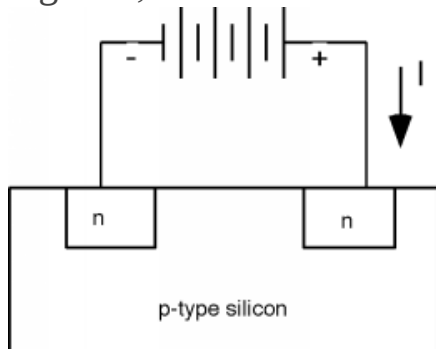
Where did these electrons come from? We do not have any donors in this material, so they can not come from there. The only place from which electrons could be found would be through thermal generation. Remember, in a semiconductor, there are always a few electron hole pairs being generated by thermal excitation at any given time. Electrons that get created in the depletion region are caught by the electric field and are swept over to the edge by the gate. I have tried to suggest this with the electron generation event shown in the band diagram in the figure. In a real MOS device, we have the two n-regions, and it is easy for electrons from one or both to "fall" into the potential well under the gate, and create the inversion layer of electrons.

Introduction to the MOS Transistor and MOSFETs

Note: This module is adapted from the Connexions modules entitled *Introduction to MOSFETs* and *MOS Transistor* by Bill Wilson.

We now move on to another three terminal device - also called a transistor. This transistor, however, works on much different principles than does the bipolar junction transistor of the last chapter. We will now focus on a device called the *field effect transistor*, or *metal-oxide-semiconductor field effect transistor* or simply MOSFET.

In [\[link\]](#) we have a block of silicon, doped p-type. Into it we have made two regions which are doped n-type. To each of those n-type regions we attach a wire, and connect a battery between them. If we try to get some current, I , to flow through this structure, nothing will happen, because the n-p junction on the RHS is reverse biased, i.e., the positive lead from the battery going to the n-side of the p-n junction. If we attempt to remedy this by turning the battery around, we will now have the LHS junction reverse biased, and again, no current will flow. If, for whatever reason, we want current to flow, we will need to come up with some way of forming a layer of n-type material between one n-region and the other. This will then connect them together, and we can run current in one terminal and out the other.

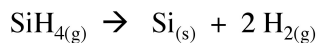


The start of a field effect

transistor.

To see how we will do this, let's do two things. First we will grow a layer of SiO₂ (silicon dioxide or silica, but actually referred to as "oxide") on top of the silicon. To do this the wafer is placed in an oven under an oxygen atmosphere, and heated to 1100 °C. The result is a nice, high-quality insulating SiO₂ layer on top of the silicon). On top of the oxide layer we then deposit a conductor, which we call the gate. In the "old days" the gate would have been a layer of aluminum; hence the "metal-oxide-silicon" or MOS name. Today, it is much more likely that a heavily doped layer of polycrystalline silicon (polysilicon, or more often just "poly") would be deposited to form the gate structure. Polysilicon is made from the reduction of a gas, such as silane (SiH₄), [\[link\]](#).

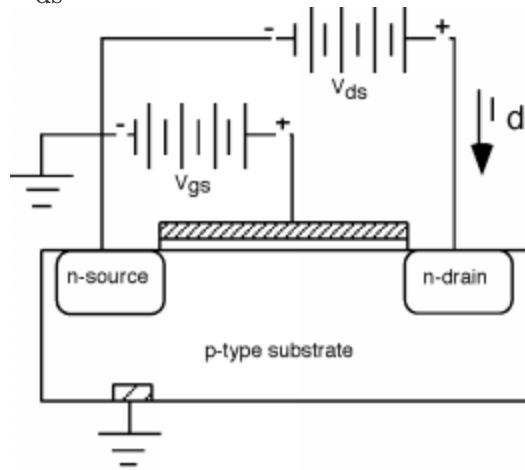
Equation:



The silicon is polycrystalline (composed of lots of small silicon crystallites) because it is deposited on top of the oxide, which is amorphous, and so it does not provide a single crystal "matrix" which would allow the silicon to organize itself into one single crystal. If we had deposited the silicon on top of a single crystal silicon wafer, we would have formed a single crystal layer of silicon called an epitaxial layer. This is sometimes done to make structures for particular applications. For instance, growing a n-type epitaxial layer on top of a p-type substrate permits the fabrication of a very abrupt p-n junction.

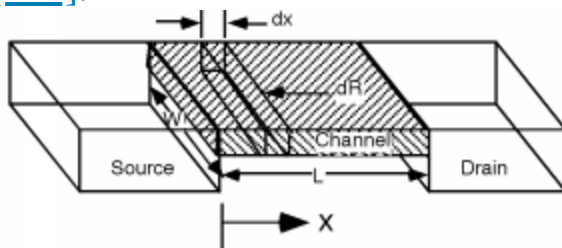
Note: Epitaxy, is a transliteration of two Greek words *epi*, meaning "upon", and *taxis*, meaning "ordered". Thus an epitaxial layer is one that follows the order of the substrate on which it is grown.

Now we can go back now to our initial structure, shown in [\[link\]](#), only this time we will add an oxide layer, a gate structure, and another battery so that we can invert the region under the gate and connect the two n-regions together. We'll also identify some names for parts of the structure, so we will know what we are talking about. For reasons which will be clear later, we call the n-region connected to the negative side of the battery the source, and the other one the drain. We will ground the source, and also the p-type substrate. We add two batteries, V_{gs} between the gate and the source, and V_{ds} between the drain and the source.



Biassing a MOSFET transistor

It will be helpful if we also make another sketch, which gives us a perspective view of the device. For this we strip off the gate and oxide, but we will imagine that we have applied a voltage greater than V_T to the gate, so there is a n-type region, called the channel which connects the two. We will assume that the channel region is L long and W wide, as shown in [\[link\]](#).



The inversion channel and its resistance.

Next we want to take a look at a little section of channel, and find its resistance $\mathcal{d}(R)$, when the little section is $\mathcal{d}(x)$ long, [\[link\]](#).

Equation:

$$\mathcal{d}(R) = \frac{dx}{\sigma_s W}$$

We have introduced a slightly different form for our resistance formula here. Normally, we would have a simple σ in the denominator, and an area A , for the cross-sectional area of the channel. It turns out to be very hard to figure out what that cross sectional area of the channel is however. The electrons which form the inversion layer crowd into a very thin sheet of surface charge which really has little or no thickness, or penetration into the substrate.

If, on the other hand we consider a surface conductivity (units: simply mhos), σ_s , [\[link\]](#), then we will have an expression which we can evaluate. Here, μ_s is a surface mobility, with units of $\text{cm}^2/\text{V}\cdot\text{sec}$, that is the quantity which represented the proportionality between the average carrier velocity and the electric field, [\[link\]](#) and [\[link\]](#).

Equation:

$$\sigma_s = \mu_s Q_{\text{chan}}$$

Equation:

$$\bar{v} = \mu E$$

Equation:

$$\mu = \frac{q\tau}{m}$$

The surface mobility is a quantity which has to be measured for a given system, and is usually just a number which is given to you. Something around $300 \text{ cm}^2/\text{V}\cdot\text{sec}$ is about right for silicon. Q_{chan} is called the surface charge density or channel charge density and it has units of Coulombs/cm². This is like a sheet of charge, which is different from the bulk charge density, which has units of Coulombs/cm³. Note that:

Equation:

$$\begin{aligned} \frac{\text{cm}^2}{\text{Volt sec}} \frac{\text{Coulombs}}{\text{cm}^2} &= \frac{\frac{\text{Coul}}{\text{sec}}}{\text{Volt}} \\ &= \frac{I}{V} \\ &= \text{mhos} \end{aligned}$$

It turns out that it is pretty simple to get an expression for Q_{chan} , the surface charge density in the channel. For any given gate voltage V_{gs} , we know that the charge density on the gate is given simply as:

Equation:

$$Q_g = c_{\text{ox}} V_{\text{gs}}$$

However, until the gate voltage V_{gs} gets larger than V_T we are not creating any mobile electrons under the gate, we are just building up a depletion region. We'll define Q_T as the charge on the gate necessary to get to threshold. $Q_T = c_{\text{ox}} V_T$. Any charge added to the gate above Q_T is matched by charge Q_{chan} in the channel. Thus, it is easy to say: [\[link\]](#) or [\[link\]](#).

Equation:

$$Q_{\text{channel}} = Q_g - Q_T$$

Equation:

$$Q_{\text{chan}} = c_{\text{ox}} (V_g - V_T)$$

Thus, putting [\[link\]](#) and [\[link\]](#) into [\[link\]](#), we get:

Equation:

$$\mathcal{d}(R) = \frac{\mathcal{d}(x)}{\mu_s c_{\text{ox}} (V_{\text{gs}} - V_T) W}$$

If you look back at [\[link\]](#), you will see that we have defined a current I_d flowing into the drain. That current flows through the channel, and hence through our little incremental resistance $\mathcal{d}(R)$, creating a voltage drop $\mathcal{d}(V_c)$ across it, where V_c is the channel voltage, [\[link\]](#).

Equation:

$$\begin{aligned} \mathcal{d}(V_c(x)) &= I_d \mathcal{d}(R) \\ &= \frac{I_d \mathcal{d}(x)}{\mu_s c_{\text{ox}} (V_{\text{gs}} - V_T) W} \end{aligned}$$

Let's move the denominator to the left, and integrate. We want to do our integral completely along the channel. The voltage on the channel $V_c(x)$ goes from 0 on the left to V_{ds} on the right. At the same time, x is going from 0 to L . Thus our limits of integration will be 0 and V_{ds} for the voltage integral $\mathcal{d}(V_c(x))$ and from 0 to L for the x integral $\mathcal{d}(x)$.

Equation:

$$\int_0^{V_{\text{ds}}} \mu_s c_{\text{ox}} (V_{\text{gs}} - V_T) W \, dV_c = \int_0^L I_d \, dx$$

Both integrals are pretty trivial. Let's swap the equation order, since we usually want I_d as a function of applied voltages.

Equation:

$$I_d L = \mu_s c_{\text{ox}} W (V_{\text{gs}} - V_T) V_{\text{ds}}$$

We now simply divide both sides by L , and we end up with an expression for the drain current I_d , in terms of the drain-source voltage, V_{ds} , the gate

voltage V_{gs} and some physical attributes of the MOS transistor.

Equation:

$$I_d = \left(\frac{\mu_s c_{\text{ox}} W}{L} (V_{\text{gs}} - V_T) \right) V_{\text{ds}}$$

Light Emitting Diode

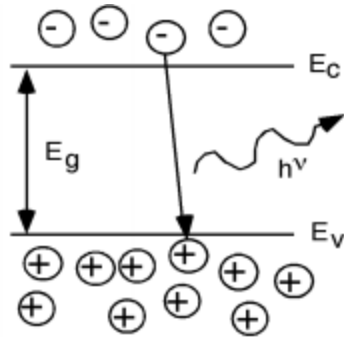
Light Emitting Diode

Note: This module is adapted from the Connexions module entitled *Light Emitting Diode* by Bill Wilson.

Let's talk about the recombining electrons for a minute. When the electron falls down from the conduction band and fills in a hole in the valence band, there is an obvious loss of energy. The question is; where does that energy go? In silicon, the answer is not very interesting. Silicon is what is known as an *indirect band-gap* material. What this means is that as an electron goes from the bottom of the conduction band to the top of the valence band, it must also undergo a significant change in momentum. This all comes about from the details of the band structure for the material, which we will not concern ourselves with here. As we all know, whenever something changes state, we must still conserve not only energy, but also momentum. In the case of an electron going from the conduction band to the valence band in silicon, both of these things can only be conserved if the transition also creates a quantized set of lattice vibrations, called *phonons*, or "heat". Phonons possess both energy and momentum, and their creation upon the recombination of an electron and hole allows for complete conservation of both energy and momentum. All of the energy which the electron gives up in going from the conduction band to the valence band (1.1 eV) ends up in phonons, which is another way of saying that the electron heats up the crystal.

In some other semiconductors, something else occurs. In a class of materials called *direct band-gap* semiconductors, the transition from conduction band to valence band involves essentially no change in momentum. Photons, it turns out, possess a fair amount of energy (several eV/photon in some cases) but they have very little momentum associated with them. Thus, for a direct band gap material, the excess energy of the electron-hole recombination can either be taken away as heat, or more likely, as a photon of light. This radiative transition then conserves energy

and momentum by giving off light whenever an electron and hole recombine. This gives rise to the light emitting diode (LED). Emission of a photon in an LED is shown schematically in [\[link\]](#).



Radiative recombination in a direct band-gap semiconductor.

It was Planck who postulated that the energy of a photon was related to its frequency by a constant, which was later named after him. If the frequency of oscillation is given by the Greek letter "nu" (ν), then the energy of the photon is just given by, [\[link\]](#), where h is Planck's constant, which has a value of 4.14×10^{-15} eV.sec.

Equation:

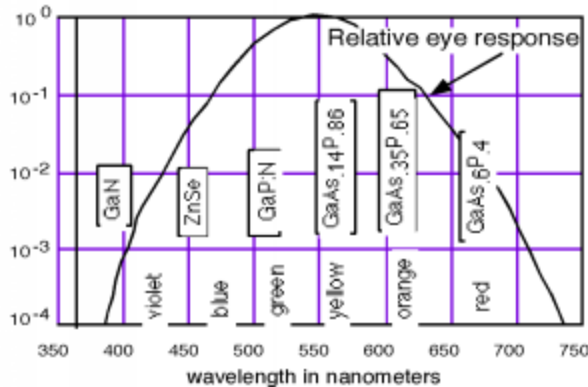
$$E = h\nu$$

When we talk about light it is conventional to specify its wavelength, λ , instead of its frequency. Visible light has a wavelength on the order of nanometers, e.g., red is about 600 nm, green about 500 nm and blue is in the 450 nm region. A handy "rule of thumb" can be derived from the fact that $c = \lambda\nu$, where c is the speed of light (3×10^3 m/sec or 3×10^{17} nm/sec, [\[link\]](#)).

Equation:

$$\begin{aligned} \lambda(\text{nm}) &= \frac{hc}{E(\text{eV})} \\ &= \frac{1242}{E(\text{eV})} \end{aligned}$$

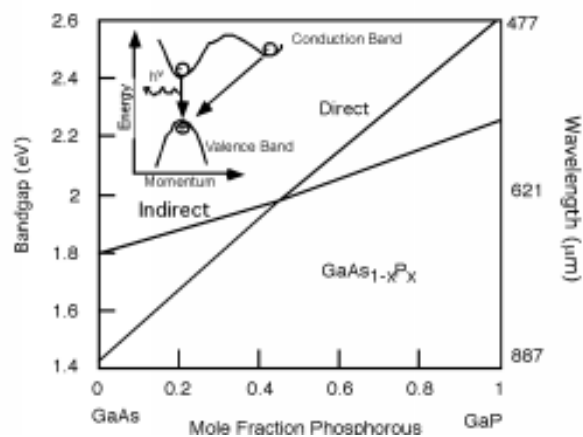
Thus, a semiconductor with a 2 eV band-gap should give off light at about 620 nm (in the red). A 3 eV band-gap material would emit at 414 nm, in the violet. The human eye, of course, is not equally responsive to all colors ([\[link\]](#)). The materials which are used for important light emitting diodes (LEDs) for each of the different spectral regions are also shown in [\[link\]](#).



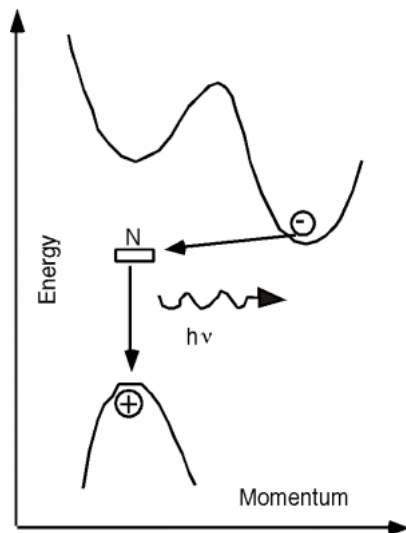
Relative response of the human eye to various colors.

It is worth noting that a number of the important LEDs are based on the GaAsP system. GaAs is a direct band-gap semiconductor with a band gap of 1.42 eV (in the infrared). GaP is an indirect band-gap material with a band gap of 2.26 eV (550 nm, or green). Both As and P are group V elements. (Hence the nomenclature of the materials as III-V (or 13-15) compound semiconductors.) We can replace some of the As with P in GaAs and make a mixed compound semiconductor $\text{GaAs}_{1-x}\text{P}_x$. When the mole fraction of phosphorous is less than about 0.45 the band gap is direct, and so we can "engineer" the desired color of LED that we want by simply growing a crystal with the proper phosphorus concentration! The properties of the GaAsP system are shown in [\[link\]](#). It turns out that for this system, there are actually two different band gaps, as shown in [\[link\]](#). One is a direct gap (no change in momentum) and the other is indirect. In GaAs, the direct gap has lower energy than the indirect one (like in the inset) and so the transition is a radiative one. As we start adding phosphorous to the system, both the direct and indirect band gaps increase in energy. However, the direct gap energy increases faster with phosphorous fraction than does

the indirect one. At a mole fraction x of about 0.45, the gap energies cross over and the material goes from being a direct gap semiconductor to an indirect gap semiconductor. At $x = 0.35$ the band gap is about 1.97 eV (630 nm), and so we would only expect to get light up to the red using the GaAsP system for making LED's. Fortunately, people discovered that you could add an impurity (nitrogen) to the GaAsP system, which introduced a new level in the system. An electron could go from the indirect conduction band (for a mixture with a mole fraction greater than 0.45) to the nitrogen site, changing its momentum, but not its energy. It could then make a direct transition to the valence band, and light with colors all the way to the green became possible. The use of a nitrogen recombination center is depicted in the [\[link\]](#).



Band gap for the GaAsP system



Addition of a
nitrogen
recombination
center to indirect
GaAsP.

If we want colors with wavelengths shorter than the green, we must abandon the GaAsP system and look for more suitable materials. A compound semiconductor made from the II-VI elements Zn and Se make up one promising system, and several research groups have successfully made blue and blue-green LEDs from ZnSe. SiC is another (weak) blue emitter which is commercially available on the market. Recently, workers at a tiny, unknown chemical company stunned the "display world" by announcing that they had successfully fabricated a blue LED using the II-V material GaN. A good blue LED was the "holy grail" of the display and CD ROM research community for a number of years. Obviously, adding blue to the already working green and red LED's completes the set of 3 primary colors necessary for a full-color flat panel display. Furthermore, using a blue LED or laser in a CD ROM would more than quadruple its data capacity, as bit diameter scales as λ , and hence the area as λ^2 .

Polymer Light Emitting Diodes

This module was developed as part of a Rice University course CHEM496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Pui Yee Hung.

Introduction

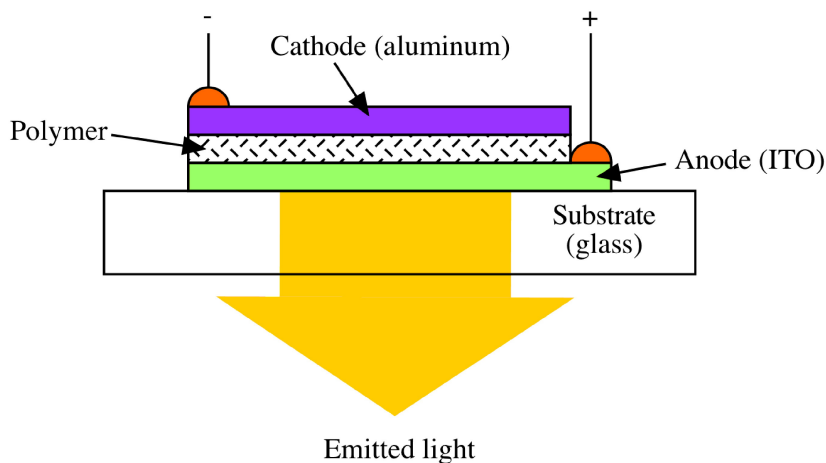
In 1990, electroluminescent (EL) from conjugated polymers was first reported by Burroughes et al. of Cambridge University. A layer of poly(*para*-phenylenevinylene) (PPV) was sandwiched between layers of indium tin oxide (ITO) and aluminum. When this device is under a 14 V dc bias, the PPV emits a yellowish-green light with a quantum efficiency of 0.05%. This report attracted a lot of attention, because the potential that polymer light emitting diodes (LEDs) could be inexpensively mass produced into large area display area. The processing steps in making polymer LEDs are readily scaleable. The industrial coating techniques is well developed to mass produce polymer layers of 100 nm thickness, and the device could be patterned onto large surface area by pixellation of metal.

Since the initial discovery, and increasing amount of researches has been performed, and significant progress has been made. In 1990 the polymer LED only emitted yellowish green color, now the emission color ranged from deep blue to near infra red. The efficiency of the multi-layer polymer LED even reached a quantum efficiency of >4% and the operating voltage has been reduced significantly. In term of efficiency, color selection and operating voltage, polymer LEDs have attained adequate levels for commercialization. But there are reliability problems that are symptomatic of any organic devices.

Device physics and materials science of polymer LEDs

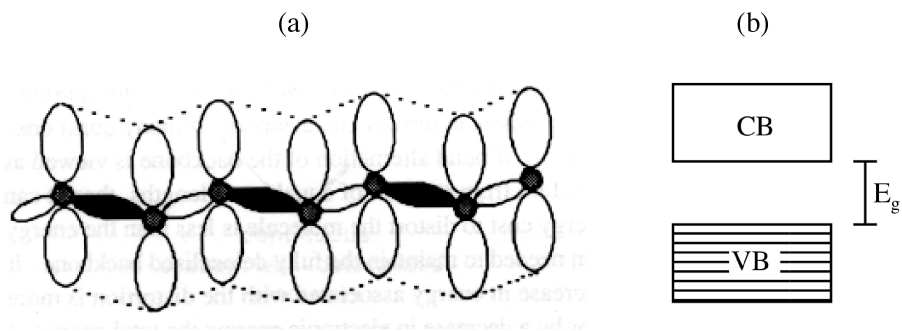
A schematic diagram of a polymer LED is shown in [\[link\]](#). A polymer LED can be divided into three different components:

- A. **Anode:** the hole supplier, made of metal of high working function. Examples of the common anode are indium tin oxide (ITO), gold etc. The anode is usually transparent so that light can be emitted through.
- B. **Cathode:** the electron supplier, made of metal of low working function. Examples of the common cathode are aluminum or calcium.
- C. **Polymer:** made of conjugated polymer film with thickness of 100 nm.

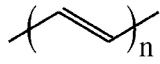
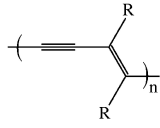
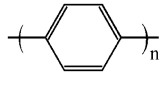


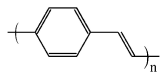
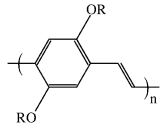
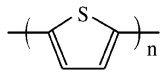
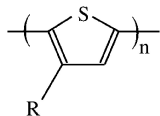
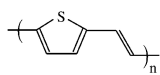
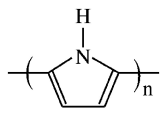
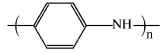
Schematic set-up of polymer LED.

When a polymer LED is under a direct current (dc) bias, holes are injected from the anode (ITO) and electrons are injected from the cathode (aluminum). Under the influences of the electrical field, the electrons and holes will migrate toward each other. When they recombine in the conjugated polymer layer, a bound excited states (excitons) will be formed. Some of the excitons (singlets) then decays in the conjugated polymer layer to emit light through the transparent substrates (glass). The emission color will be depended on the energy gap of the polymers. There is energy gap in a conjugated polymer because the π electron are not completely delocalized over the entire polymer chain. Instead there are alternate region in the polymer chain that has a higher electron density ([\[link\]a](#)). The chain length of this region is about 15-20 multiple bonds. The emission color can be controlled by tuning this energy band gap ([\[link\]b](#)). It shows that bond alternation limits the extent of delocalization. [\[link\]](#) summarizes the structure and emission color of some common conjugated polymers.



Alternation of bond lengths along a conjugated polymer chain (a) results in a material with properties of a large band gap semiconductor (b), where CB is the conductive band gap, and VB is the valence band, and E_g is the band gap.

Polymer	Chemical name	Structure	π - π^* energy gap (eV)	Emission peak (nm)
PA	<i>trans</i> -polyacetylene		1.5	600
PDA	polydiacetylene		1.7	
PPP	poly(<i>para</i> -phenylene)		3.0 (red)	465

PPV	Poly(<i>para</i> -phenylenevinylene)		2.5 (green)	565
RO-PPV	poly(2,5-dialkoxy-p-phenylenevinylene)		2.2 (blue)	~580
PT	polythiophene		2.0 (red)	
P3AT	Poly(3-alkylthiophene)		2.0 (red)	690
PTV	Poly(2,5-thiophenevinylene)		1.8	
PPy	Polypyrrole		3.1	
PAni	Polyaniline			3.2

Example of common conjugated polymers.

Approaches to improve the efficiency

Efficiency for any LED is defined:

$$\eta_{\text{ext}} = \eta_{\text{esc}} * \eta_{\text{int}}$$

where n_{ext} is the external quantum efficiency, n_{int} is the internal efficiency (represents the fraction of injected carrier, usually electron, that is converted to photon), and n_{esc} is the escape efficiency (represent fraction of photons that can reach to the outside).

The most common way to improve the internal efficiency is to balance the number of electrons and holes which arrives at the polymer layer. Originally, there are more holes than electron that arrive of the polymer layer because conjugated polymers have a higher electron affinity, and as a consequence will favor the transport of hole than electron. There are two ways to maintains the balance:

1. Match the work function of electrode with the electron affinity and ionization potential of the polymer.
2. Tune the polymer's electron affinity and ionization potential to match the work function of the electrode.

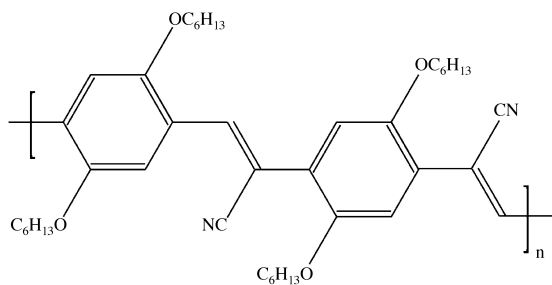
The escape efficiency is also important because a polymer LED is made up of layers of materials that have different refractive index, and some of the photon generated from the excitation may be reflected at the boundary and trapped inside the device.

Improvement in internal quantum efficiency using low working function cathode

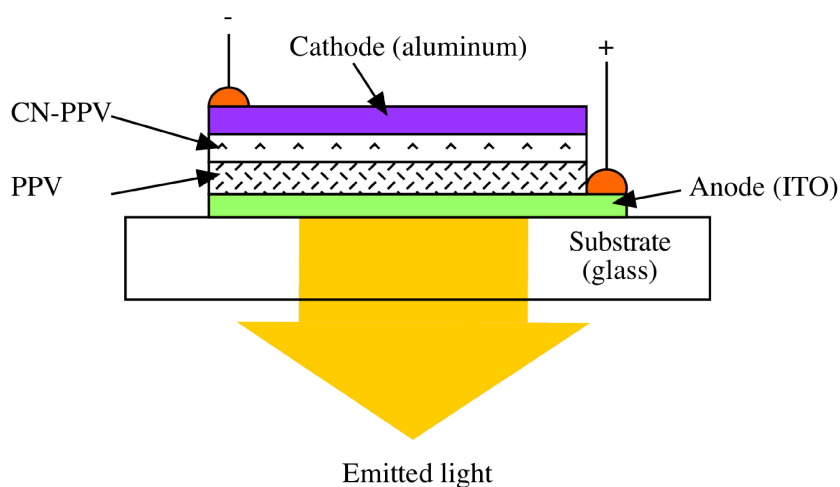
Conjugated polymer is electron rich, the mobility for hole is higher than electron, and more holes will arrive in the polymer layer than electrons. One way to increase the population of the electron is to use a lower working function metal as cathode. Braun and Heeger have replaced the aluminum cathode with calcium results in improved internal efficiency by a factor of ten, to 0.1%. This approach is direct and fast but low working function electrode like calcium will be oxidized easily and shorten the devices' life.

Improvement in internal quantum efficiency using multiple polymer layers

A layer of poly[2,5-di(hexyloxy)cyanoterephthalylidene] (CN-PPV, [\[link\]](#)) is coated on top of PPV to improve the transport and recombination of electron and holes ([\[link\]](#)).



Structure of CN-PPV.



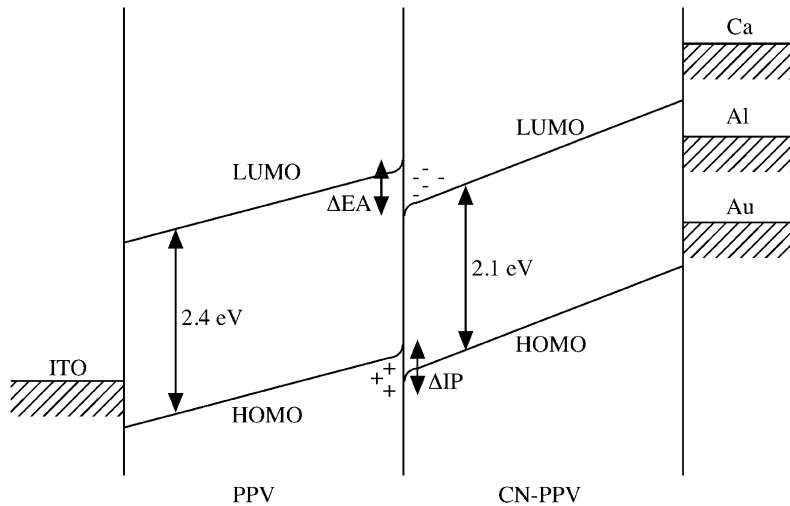
Schematic representation of a CN-PPV and PPV multi-layer LED.

The nitrile group in the CN-PPV has two effect on the polymer.

1. It increases the electron affinity so electrons can travel more efficient from the aluminum to the polymer layer. And metal of relative high working function like aluminum and gold can be now be used as cathode instead of calcium.
2. It increases the binding energy of the occupied π and unoccupied π^* state but maintain a similar π - π^* gap. So when the PPV and CN-PPV is placed

together, holes and electron will be confined at the heterojunction.

The resulting energy levels are shown in [\[link\]](#).



Schematic energy-level diagram for a PPV and CN-PPV under forward bias. Adapted from N. C. Greenham, S. C. Maratti, D. D. C. Bradley, R. H. Friend, and A. B. Holmes, *Nature*, 1993, **365**, 62.

The absolute energies of levels are not known accurately, but the diagram shows the relative position of the HOMO and LUMO levels in the polymers, and the Fermi levels of the various possible metal contacts, the differences in electron affinity (ΔEA) and ionization potential (ΔIP) between PPV and CN-PPV are also shown ([\[link\]](#)).

At the polymers interface there is a sizable offset in the energies of HOMO and LUMO of PPV and CN-PPV, the holes transported from the ITO and the electrons transport from the aluminum will be confined in the heterojunction. The local charge density will be sufficiently high to ensure the holes and electrons will pass within a collision capture radius. This set-up increases the chance for an electrons to combine with holes to form an excitation. In addition, the emission will be close to the junction, far away from the electrode junction which will quench the singlet excitations. The result is that a multi-layers LED has an internal quantum efficiency

of 10% and external quantum efficiency (for light emitted in forward direction) of 25%.

Based on this approach, a couple of polymers have been developed or modified to produce the desirable emission color and processing property. The drawback of this method is that desirable properties may not be complementary to each other. For example, in MEH-PPV an alkoxy side group (RO) is introduced to PPV so that it can be dissolved in organic solvent. But the undesirable effect is that MEH-PPV is less thermally stable. Moreover in multiple layers LEDs, different polymer layers have different refractive indices and a fraction of the photons will undergo total internal reflection at the refractive boundaries and cannot escape as light. This problem can be overcome by Fabry-Pert microcavity structure.

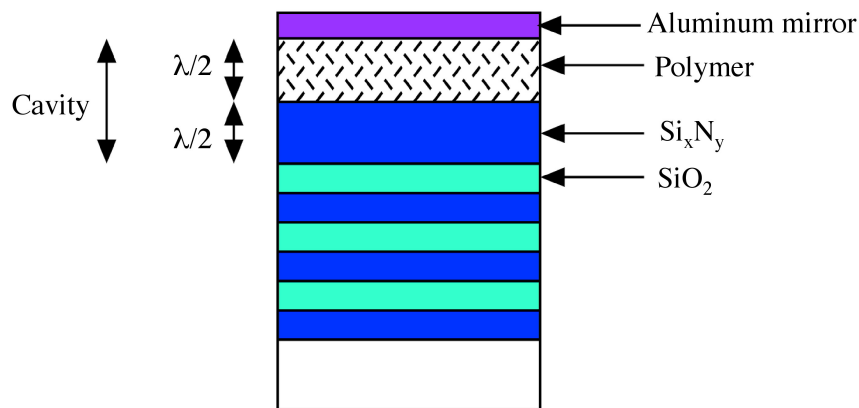
Improvement in external quantum efficiency using microcavity

Fabry-Perot resonant structures are also used in inorganic LED, and are based on Fermi's golden rule:

$$K_r \sim | \langle M \rangle |^2 r_{(v)}$$

where M (the matrix element of the perturbation between final and initial states) depends on the nature of the material, and $r_{(v)}$ can be altered by changing the density of various density states, e.g. using a luminescent thin films to select certain value of V.

In building a microcavity for a polymer LED, the polymer is placed between two mirrors. ([link](#)), in which one of the mirrors is made up of aluminum, the other mirror (a Bragg Mirror) is formed by epitaxial multilayer stacks of Si_xN_y and SiO_2 .



Schematic set-up of micro-cavity.

Improvement in internal quantum efficiency: doping of polymer

Doping is a process that creates carrier by purposely introducing impurities and is very popular method in the semiconductor industry. However, this technique was not used in polymer LED until 1995, when a co-polymer polystyrene-poly(3-hexylthiophene) (PS-P3HT) was doped with FeCl_3 . Doping of MEH-PPV with iodine has improved the efficiency by 200% and the polymer LED can be operated under both forward and reverse bias ([\[link\]](#)). The doping is accomplished by mixing 1 wt% MEH-PPV with 0.2 wt% I_2 . The molar ratio of MEH-PPV to I_2 is 5:1. That is a huge “doping “ ratio when you compare the doping concentration in the semiconductor.

	Un-doped	Doped
Turn on voltage (V)	10	foreword 5, reversed 12
External efficiency (%)	4×10^{-4}	8×10^{-3}

Results of iodine doping of an Al/MEH-PPV/ITO-based LED.

Polymer LEDs on a silicon substrate: an application advantage over inorganic LEDs

In the initial research polymer LEDs were in direct competition with the inorganic LEDs and tried to achieve the existing LED standard. This is a difficult task as polymer LEDs have a lower long term stability. However, there are some applications in which polymer LEDs have a clear advantage over their more traditional inorganic analogs. One of these is to incorporate LEDs with the silicon integrated circuits for inter-chip communication.

It is difficult to build inorganic LEDs on a silicon substrate, because of the thermal stress developing between the inorganic LED (usually a III-V based device) and the silicon interface. But polymer LEDs offer a solution, since polymers can be easily spin-coated on the silicon. The operating voltage of polymer LED is less than 4 V, and the turn on voltage can be as low as 2 V. Together with a switching time of less than 50 ns, make polymer LED a perfect candidate.

Reliability and degradation of polymer LEDs

In terms of the efficiency, color selection, and driving voltage, polymer LED have attained adequate level for commercialization. However, the device lifetime is still far from satisfactory. Research into understanding the reliability and degradation mechanisms of polymer LEDs has generally been divided into two area:

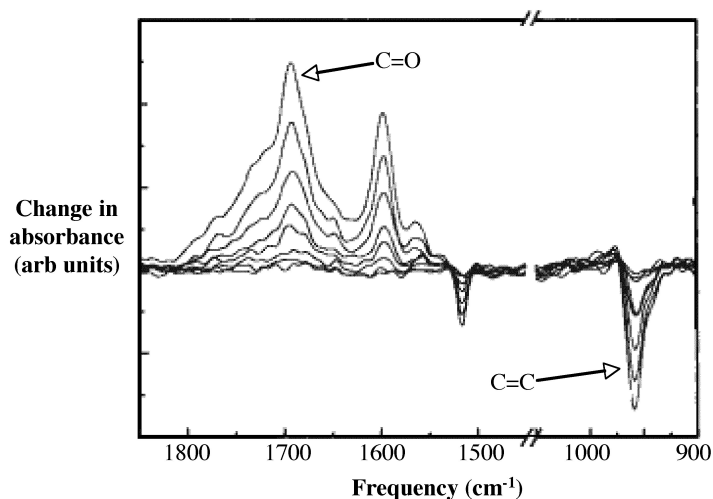
1. Photo-degradation of polymer.
2. Interface degradation.

Polymer photo degradation

Photoluminescence (PL) studies of the photo-oxidation of PPV have been undertaken, since it is believed that EL is closely related with PL.

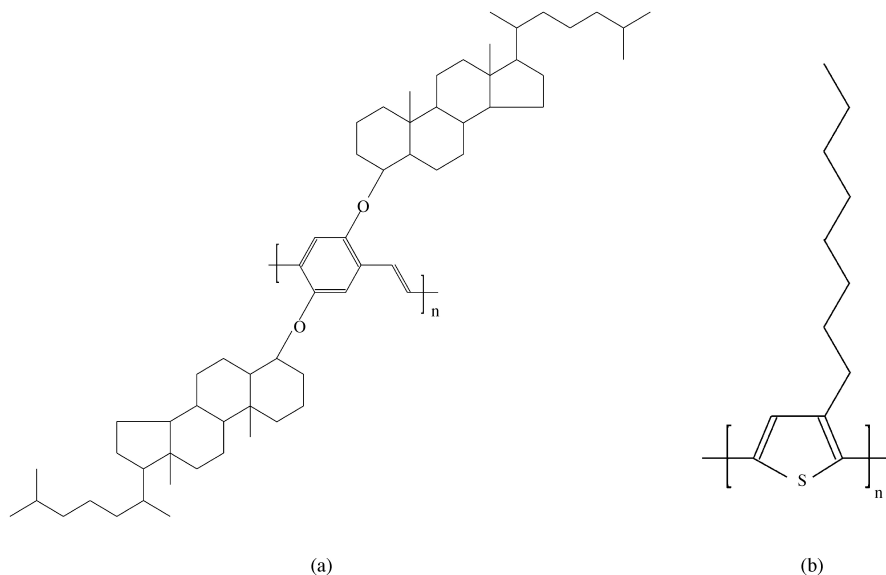
It was found that there is a rapid decay in emission when PPV is exposed to oxygen. Using time resolved FTIR spectroscopy an increase in the carbonyl signal and a decrease in C=C signal with time ([\[link\]](#)). It was suggested that the carbonyl

group has a strong electron affinity level to charge transfer between molecules segment in the polymer, thereby dissociating the excitation and quenching the PL.



FTIR as a function of photo-oxidation of PPV. Adapted from M. Yan, L. J. Rothberg, F. Papadimitrakopoulos, M. E. Galvin and T. M. Miller, *Phys. Rev. Lett.*, 1994, **73**, 744.

Similar research was performed by Cumpston and Jensen using BCHA-PPV and P3OT ([\[link\]](#)) and exposing them to dry air in UV irradiation. In BCHA-PPV, there is an increase in carbonyl signal with time, while the P3OT remain intact. A mechanism proposed for the degradation of BCHA-PPV involves the transfer of energy from the excited triplet state of the PPV to oxygen to form singlet oxygen which attacks the vinyl double bond in the PPV backbone. And P3OT does not have a vinyl bond so it can resist the oxidation.



Structure of (a) BCHA-PPV and (b) P3OT.

The research described above was all performed on polymer thin films deposited on an inert surface. The presence of cathode and anode may also affect the oxidation mechanism. Scott et al. have taken IR spectra from a MEH-PPV LED in the absence of oxygen. They obtained similar result as in Yan et al., however, a decrease in ITO's oxygen signal was noticed suggesting that the ITO anode acts like a oxygen reservoir and supplies the oxygen for the degradation process.

Polymer LED interface degradation

There are few interface degradation studies in polymer LEDs. One of them by Scott et al. took SEM image of the cathode from a failed polymer LED. The polymer LED used ITO as the anode, MEH-PPV as the polymer layer, and an aluminum calcium alloy as cathode. SEM images showed “craters” formed in the cathode. The craters are formed when the cathode metal is melted and pull away from the polymer layer. It was suggested that a high current density will generate heat and result in local hot spot. The temperature in the hot spot is high enough to melt the cathode. And when it melt, it will pull away from the polymer. This process will decrease the effective cathode area, and reduce the luminescence gradually.

Bibliography

- D. R. Baigent, N. C. Greenham, J. Gruner, R. N. Marks, R. H. Friend, S. C. Moratti, and A. B. Holmes, *Synth.Met.*, 1994, **67**, 3.
- B. H. Cumpston and K. F. Jensen, *Synth. Met.*, 1995, **73**, 195.
- J. H. Burroughes, D. D. C. Bradley, A. R. Brown, R. N. Marks, K. Mackay, R. H. Friend, P. L. Burns, and A. B. Holmes, *Nature*, 1990, **347**, 539.
- N. C. Greenham, S. C. Maratti, D. D. C. Bradley, R. H. Friend, and A. B. Holmes, *Nature*, 1993, **365**, 628.
- J. Gruner, F. Cacialli, I. D. W. Samuel, R. H. Friend, *Synth. Met*, 1996, **76**, 197.
- M. Herold, J. Gmeiner, W. Riess, and M. Schwoerer, *Synth. Met.*, 1996, **76**, 109.
- R. H. Jordan, A. Dodabalapur, L. J. Rothberg, and R. E. Slusher, *Proceeding of SPIE*, 1997, **3002**, 92.
- I. D. Parker and H. H. Kim, *Appl. Phys. Lett.*, 1994, **64**, 1774.
- J. C. Scott, J. Kaufman, P. J. Brock, R. DiPietro, J. Salem, and J. A. Goitia, *J. Appl. Phys.*, 1996, **79**, 2745.
- M. S. Weaver, D. G. Lidzaey, T. A. Fisher, M. A. Pate, D. O'Brien, A. Bleyer, A. Tajbakhsh, D. D. C. Bradley, M. S. Skolnick, and G. Hill, *Thin solid Films*, 1996, **273**, 39.
- M. Yan, L. J. Rothberg, F. Papadimitrakopoulos, M. E. Galvin, and T. M. Miller, *Phys. Rev. Lett.*, 1994, **73**, 744.

Laser

LASER

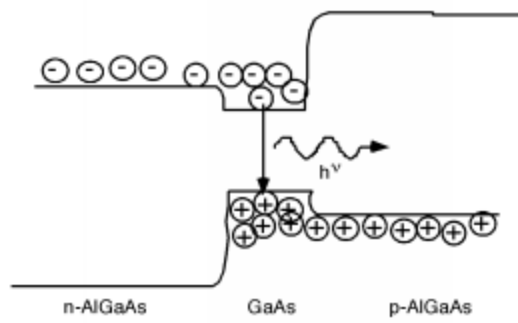
Note: This module is adapted from the Connexions module entitled *LASER* by Bill Wilson.

What is the difference between an LED and a solid state laser? There are some differences, but both devices operate on the same principle of having excess electrons in the conduction band of a semiconductor, and arranging it so that the electrons recombine with holes in a radiative fashion, giving off light in the process. What is different about a laser? In an LED, the electrons recombine in a random and unorganized manner. They give off light by what is known as *spontaneous emission*, which simply means that the exact time and place where a photon comes out of the device is up to each individual electron, and things happen in a random way.

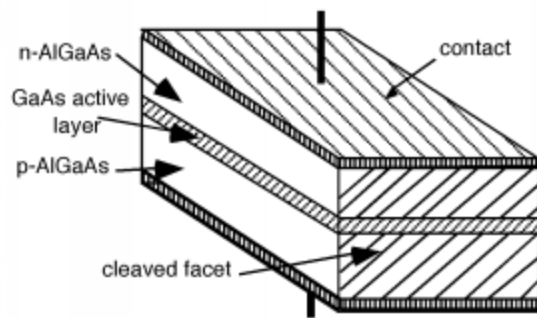
There is another way in which an excited electron can emit a photon however. If a field of light (or a set of photons) happens to be passing by an electron in a high energy state, that light field can induce the electron to emit an additional photon through a process called stimulated emission. The photon field *stimulates* the electron to emit its energy as an additional photon, which comes out *in phase with the stimulating field*. This is the big difference between *incoherent light* (what comes from an LED or a flashlight) and *coherent light* which comes from a laser. With coherent light, all of the electric fields associated with each photon are all exactly in phase. This coherence is what enables us to keep a laser beam in tight focus, and to allow it to travel a large distance without much divergence or spreading out.

So how do we restructure an LED so that the light is generated by stimulated emission rather than spontaneous emission? Firstly, we build what is called a heterostructure. All this means is that we build up a sandwich of somewhat different materials, with different characteristics. In this case, we put two wide band-gap regions around a region with a

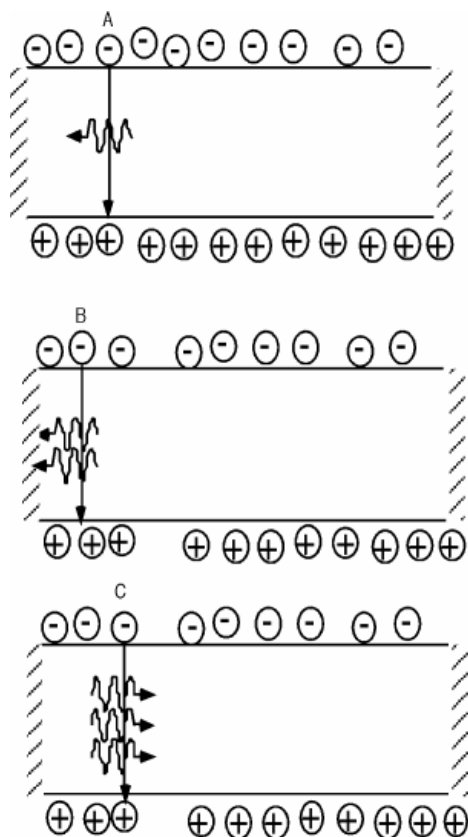
narrower band gap. The most important system where this is done is the AlGaAs/GaAs system. A band diagram for such a set up is shown in [\[link\]](#). AlGaAs (pronounced "Al-Gas") has a larger band-gap than does GaAs. The potential "well" formed by the GaAs means that the electrons and holes will be confined there, and all of the recombination will occur in a very narrow strip. This greatly increases the chances that the carriers can interact, but we still need some way for the photons to behave in the proper manner. [\[link\]](#) is a diagram of what a typical diode might look like. We have the active GaAs layer sandwiched in-between the two heterostructure confinement layers, with a contact on top and bottom. On either end of the device, the crystal has been "cleaved" or broken along a crystal lattice plane. This results in a shiny "mirror-like" surface, which will reflect photons. The back surface (which we can not see here) is also cleaved to make a mirror surface. The other surfaces are purposely roughened so that they do not reflect light. Now let us look at the device from the side, and draw just the band diagram for the GaAs region ([\[link\]](#)). We start things off with an electron and hole recombining spontaneously. This emits a photon which heads towards one of the mirrors. As the photon goes by other electrons, however, it may cause one of them to decay by stimulated emission. The two (in phase) photons hit the mirror and are reflected and start back the other way. As they pass additional electrons, they stimulate them into a transition as well, and the optical field within the laser starts to build up. After a bit, the photons get down to the other end of the cavity. The cleaved facet, while it acts like a mirror, is not a perfect one. Some light is not reflected, but rather "leaks"; though, and so becomes the output beam from the laser. The details of finding what the ratio of reflected to transmitted light is will have to wait until later in the course when we talk about dielectric interfaces. The rest of the photons are reflected back into the cavity and continue to stimulate emission from the electrons which continue to enter the gain region because of the forward bias on the diode.



The band diagram for a double heterostructure GaAs/AlGaAs laser.

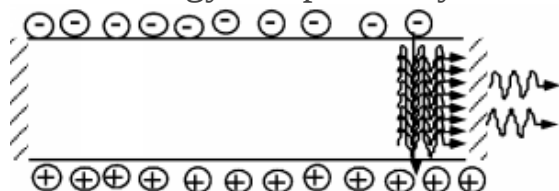


A schematic diagram of a typical laser diode.



Build up of a photon field in a laser diode.

In reality, the photons do not move back and forth in a big "clump" as we have described here, rather they are distributed uniformly along the gain region ([\[link\]](#)). The field within the cavity will build up to the point where the loss of energy by light leaking out of the mirrors just equals the rate at which energy is replaced by the recombining electrons.

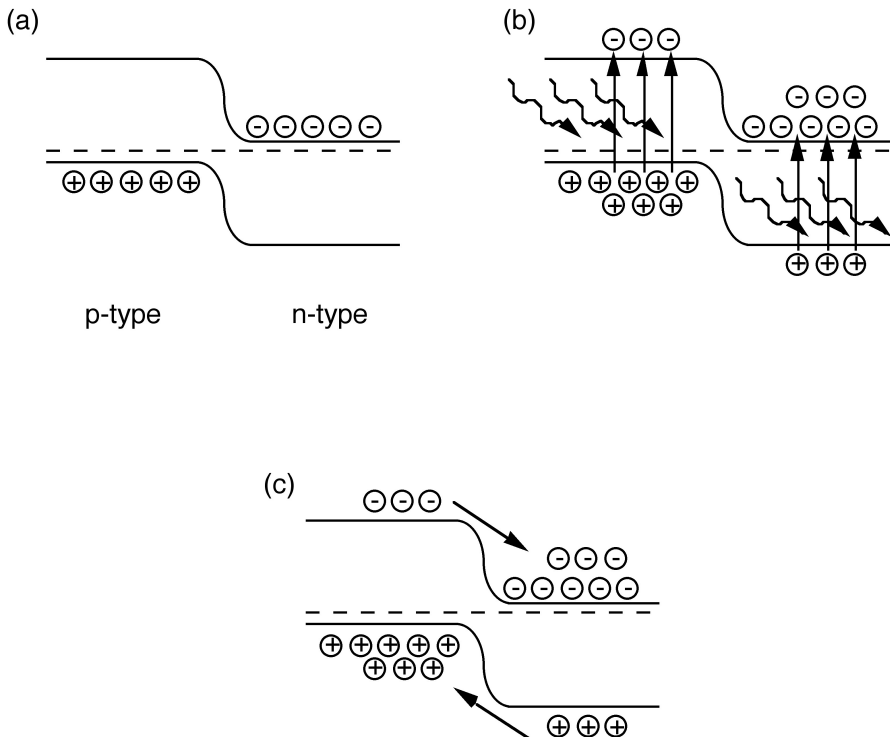


Output coupling in a diode laser.

Solar Cells

Note: This module is adapted from the Connexions module entitled *Solar Cells* by Bill Wilson.

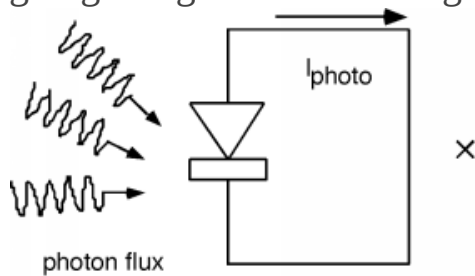
Now let us look at the opposite process of light generation for a moment. Consider the following situation where we have just a plain old normal p-n junction, only now, instead of applying an external voltage, we imagine that the junction is being illuminated with light whose photon energy is greater than the band-gap ([\[link\]](#)a). In this situation, instead of recombination, we will get photo-generation of electron hole pairs. The photons simply excite electrons from the full states in the valence band, and "kick" them up into the conduction band, leaving a hole behind. This is similar to the thermal excitation process. As can be seen from [\[link\]](#)b, this creates excess electrons in the conduction band in the p-side of the diode, and excess holes in the valence band of the n-side. These carriers can diffuse over to the junction, where they will be swept across by the built-in electric field in the depletion region. If we were to connect the two sides of the diode together with a wire, a current would flow through that wire as a result of the electrons and holes which move across the junction.



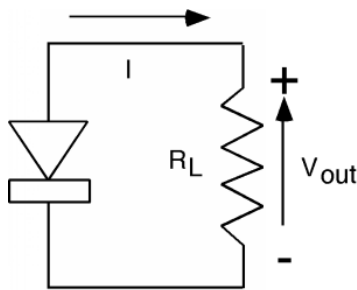
A schematic representation of a p-n diode under illumination.

Which way would the current flow? A quick look at [\[link\]](#)c shows that holes (positive charge carriers) generated on the n-side will float up to the p-side as they go across the junction. Hence positive current must be coming out of the anode, or p-side of the junction. Likewise, electrons generated on the p-side will fall down the junction potential, and come out the n-side, but since they have negative charge, this flow represents current going into the cathode. We have constructed a *photovoltaic diode*, or *solar cell*. [\[link\]](#) is a picture of what this would look like schematically. We might like to consider the possibility of using this device as a source of energy, but the way we have things set up now, since the voltage across the diode is zero, and since power equals current times voltage, we see that we are getting nada from the cell. What we need, obviously, is a load resistor, so let's put one in. It should be clear from [\[link\]](#) that the photo current flowing through the load resistor will develop a voltage which it biases the diode in the forward direction, which, of course will cause current to flow back into

the anode. This complicates things, it seems we have current coming out of the diode and current going into the diode all at the same time! How are we going to figure out what is going on?



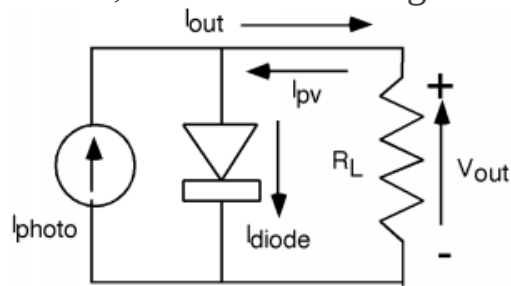
Schematic representation
of a photovoltaic cell.



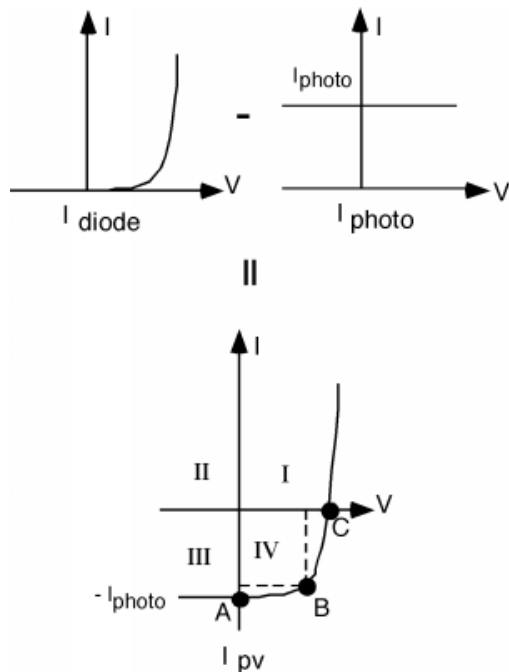
Photovoltaic cell
with a load
resistor.

The answer is to make a model. The current which arises due to the photon flux can be conveniently represented as a current source. We can leave the diode as a diode, and we have the circuit shown in [\[link\]](#). Even though we show I_{out} coming out of the device, we know by the usual polarity convention that when we define V_{out} as being positive at the top, then we should show the current for the photovoltaic, I_{pv} as current going into the top, which is what was done in [\[link\]](#). Note that $I_{pv} = I_{diode} - I_{photo}$, so all we need to do is to subtract the two currents; we do this graphically in [\[link\]](#).

Note that we have numbered the four quadrants in the I-V plot of the total PV current. In quadrant I and III, the product of I and V is a positive number, meaning that power is being dissipated in the cell. For quadrant II and IV, the product of I and V is negative, and so we are getting power from the device. Clearly we want to operate in quadrant IV. In fact, without the addition of an external battery or current source, the circuit, will only run in the IV'th quadrant. Consider adjusting R_L , the load resistor from 0 (a short) to ∞ (an open). With R_L , we would be at point A on [\[link\]](#). As R_L starts to increase from zero, the voltage across both the diode and the resistor will start to increase also, and we will move to point B, say. As R_L gets bigger and bigger, we keep moving along the curve until, at point C, where R_L is an open and we have the maximum voltage across the device, but, of course, no current coming out!



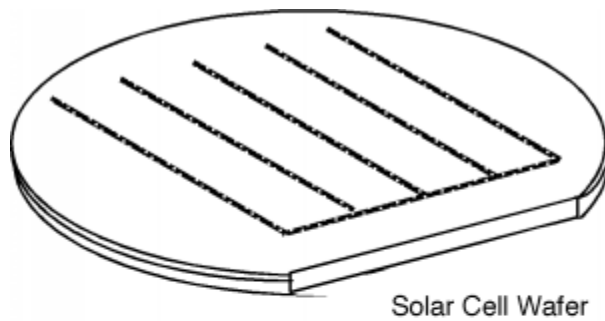
A model of a PV cell.



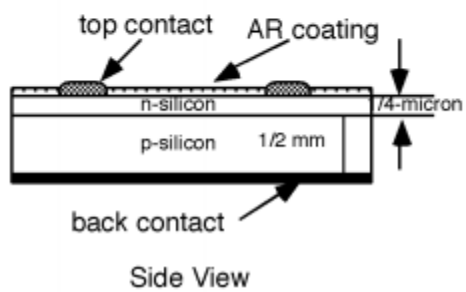
Combining the diode and
the current source.

Power is VI so at B for instance, the power coming out would be represented by the area enclosed by the two dotted lines and the coordinate axes. Somewhere about where I have point B would be where we would be getting the most power out of our solar cell.

[\[link\]](#) shows you what a real solar cell would look like. They are usually made from a complete wafer of silicon, to maximize the usable area. A shallow ($0.25 \mu\text{m}$) junction is made on the top, and top contacts are applied as stripes of metal conductor as shown. An anti-reflection (AR) coating is applied on top of that, which accounts for the bluish color which a typical solar cell has ([\[link\]](#)).



Solar Cell Wafer



Side View

A schematic diagram of a real solar cell.



A solar cell showing the blue tint due to the AR coating.

The solar power flux on the earth's surface is (conveniently) about 1 kW/m^2 or 100 mW/cm^2 . So if we made a solar cell from a 4 inch diameter wafer (typical) it would have an area of about 81 cm^2 and so would be receiving a flux of about 8.1 Watts. Typical cell efficiencies run from about 10% to maybe 15% unless special (and costly) tricks are made. This means that we will get about 1.2 Watts out from a single wafer. Looking at B on 2.59 we could guess that V_{out} will be about 0.5 to 0.6 volts, thus we could expect to get maybe around 2.5 amps from a 4 inch wafer at 0.5 volts with 15% efficiency under the illumination of one sun.

Properties of Gallium Arsenide

Gallium: the element

The element gallium was predicted, as eka-aluminum, by Mendeleev in 1870, and subsequently discovered by Lecoq de Boisbaudran in 1875; in fact de Boisbaudran had been searching for the missing element for some years, based on his own independent theory. The first experimental indication of gallium came with the observation of two new violet lines in the spark spectrum of a sample deposited on zinc. Within a month of these initial results de Boisbaudran had isolated 1 g of the metal starting from several hundred kilograms of crude zinc blende ore. The new element was named in honor of France (Latin *Gallia*), and the striking similarity of its physical and chemical properties to those predicted by Mendeleev ([\[link\]](#)) did much to establish the general acceptance of the periodic Law; indeed, when de Boisbaudran first stated that the density of Ga was 4.7 g/cm³ rather than the predicted 5.9 g/cm³, Mendeleev wrote to him suggesting that he redetermine the value (the correct value is 5.904 g/cm³).

Property	Mendeleev's prediction (1871) for eka-aluminum, M	Observed properties of gallium (discovered 1875)
Atomic weight	ca. 68	69.72
Density, g.cm ⁻³	5.9	5.904
Melting point	Low	29.78

Vapor pressure	Non-volatile	10^{-3} mmHg, 1000 °C
Valence	3	3
Oxide	M_2O_3	Ga_2O_3
Density of oxide (g/cm ³)	5.5	5.88
Properties of metal	M should dissolve slowly in acids and alkalis and be stable in air	Ga metal dissolves slowly in acids and alkalis and is stable in air
Properties of hydroxide	$M(OH)_3$ should dissolve in both acids and alkalis	$Ga(OH)_3$ dissolves in both acids and alkalis
Properties of salts	M salts will tend to form basic salts; the sulfate should form alums; M_2S_3 should be precipitated by H_2S or $(NH_4)_2S$; anhydrous MCl_3 should be more volatile than $ZnCl_2$	Ga salts readily hydrolyze and form basic salts; alums are known; Ga_2S_3 can be precipitated under special conditions by H_2S or $(NH_4)_2S$, anhydrous $GaCl_3$ is more volatile than $ZnCl_2$.

Comparison of predicted and observed properties of gallium.

Gallium has a beautiful silvery blue appearance; it wets glass, porcelain, and most other surfaces (except quartz, graphite, and Teflon[®]) and forms a brilliant mirror when painted on to glass. The atomic radius and first ionization potential of gallium are almost identical with those of aluminum and the two elements frequently resemble each other in chemical properties. Both are amphoteric, but gallium is less electropositive as indicated by its

lower electrode potential. Differences in the chemistry of the two elements can be related to the presence of a filled set of 3d orbitals in gallium.

Gallium is very much less abundant than aluminum and tends to occur at low concentrations in sulfide minerals rather than as oxides, although gallium is also found associated with aluminum in bauxite. The main source of gallium is as a by-product of aluminum refining. At 19 ppm of the earth's crust, gallium is about as abundant as nitrogen, lithium and lead; it is twice as abundant as boron (9 ppm), but is more difficult to extract due to the lack of any major gallium-containing ore. Gallium always occurs in association either with zinc or germanium, its neighbors in the periodic table, or with aluminum in the same group. Thus, the highest concentrations (0.1 - 1%) are in the rare mineral germanite (a complex sulfide of Zn, Cu, Ge, and As); concentrations in sphalerite (ZnS), bauxite, or coal, are a hundred-fold less.

Gallium pnictides

Gallium's main use is in semiconductor technology. For example, GaAs and related compounds can convert electricity directly into coherent light (laser diodes) and is employed in electroluminescent light-emitting diodes (LED's); it is also used for doping other semiconductors and in solid-state devices such as heterojunction bipolar transistors (HBTs) and high power high speed metal semiconductor field effect transistors (MESFETs). The compound MgGa_2O_4 is used in ultraviolet-activated powders as a brilliant green phosphor used in Xerox copying machines. Minor uses are as high-temperature liquid seals, manometric fluids and heat-transfer media, and for low-temperature solders.

Undoubtedly the binary compounds of gallium with the most industrial interest are those of the Group 15 (V) elements, GaE ($\text{E} = \text{N}, \text{P}, \text{As}, \text{Sb}$). The compounds which gallium forms with nitrogen, phosphorus, arsenic, and antimony are isoelectronic with the Group 14 elements. There has been considerable interest, particularly in the physical properties of these compounds, since 1952 when Welker first showed that they had semiconducting properties analogous to those of silicon and germanium.

Gallium phosphide, arsenide, and antimonide can all be prepared by direct reaction of the elements; this is normally done in sealed silica tubes or in a graphite crucible under hydrogen. Phase diagram data is hard to obtain in the gallium-phosphorus system because of loss of phosphorus from the bulk material at elevated temperatures. Thus, GaP has a vapor pressure of more than 13.5 atm at its melting point; as compared to 0.89 atm for GaAs. The physical properties of these three compounds are compared with those of the nitride in [\[link\]](#). All three adopt the zinc blende crystal structure and are more highly conducting than gallium nitride.

Property	GaN	GaP	GaAs	GaSb
Melting point (°C)	> 1250 (dec)	1350	1240	712
Density (g/cm ³)	ca. 6.1	4.138	5.3176	5.6137
Crystal structure	Würtzite	zinc blende	zinc blende	zinc blende
Cell dimen. (Å) ^a	$a = 3.187, c = 5.186$	$a = 5.4505$	$a = 5.6532$	$a = 6.0959$
Refractive index ^b	2.35	3.178	3.666	4.388
k (ohm ⁻¹ cm ⁻¹)	$10^{-9} - 10^{-7}$	$10^{-2} - 10^2$	$10^{-6} - 10^3$	6 - 13
Band gap (eV) ^c	3.44	2.24	1.424	0.71

Physical properties of 13-15 compound semiconductors. ^a Values given for 300 K. ^b Dependent on photon energy; values given for 1.5 eV incident photons. ^c Dependent on temperature; values given for 300 K.

Gallium arsenide versus silicon

Gallium arsenide is a compound semiconductor with a combination of physical properties that has made it an attractive candidate for many electronic applications. From a comparison of various physical and electronic properties of GaAs with those of Si ([\[link\]](#)) the advantages of GaAs over Si can be readily ascertained. Unfortunately, the many desirable properties of gallium arsenide are offset to a great extent by a number of undesirable properties, which have limited the applications of GaAs based devices to date.

Properties	GaAs	Si
Formula weight	144.63	28.09
Crystal structure	zinc blende	diamond
Lattice constant	5.6532	5.43095
Melting point (°C)	1238	1415
Density (g/cm ³)	5.32	2.328
Thermal conductivity (W/cm.K)	0.46	1.5
Band gap (eV) at 300 K	1.424	1.12
Intrinsic carrier conc. (cm ⁻³)	1.79 x 10 ⁶	1.45 x 10 ¹⁰

Intrinsic resistivity (ohm.cm)	10^8	2.3×10^5
Breakdown field (V/cm)	4×10^5	3×10^5
Minority carrier lifetime (s)	10^{-8}	2.5×10^{-3}
Mobility ($\text{cm}^2/\text{V.s}$)	8500	1500

Comparison of physical and semiconductor properties of GaAs and Si.

Band gap

The band gap of GaAs is 1.42 eV; resulting in photon emission in the infra-red range. Alloying GaAs with Al to give $\text{Al}_x\text{Ga}_{1-x}\text{As}$ can extend the band gap into the visible red range. Unlike Si, the band gap of GaAs is direct, i.e., the transition between the valence band maximum and conduction band minimum involves no momentum change and hence does not require a collaborative particle interaction to occur. Photon generation by inter-band radiative recombination is therefore possible in GaAs. Whereas in Si, with an indirect band-gap, this process is too inefficient to be of use. The ability to convert electrical energy into light forms the basis of the use of GaAs, and its alloys, in optoelectronics; for example in light emitting diodes (LEDs), solid state lasers (light amplification by the stimulated emission of radiation).

A significant drawback of small band gap semiconductors, such as Si, is that electrons may be thermally promoted from the valence band to the conduction band. Thus, with increasing temperature the thermal generation of carriers eventually becomes dominant over the intentionally doped level of carriers. The wider band gap of GaAs gives it the ability to remain 'intentionally' semiconducting at higher temperatures; GaAs devices are generally more stable to high temperatures than a similar Si devices.

Carrier density

The low intrinsic carrier density of GaAs in a pure (undoped) form indicates that GaAs is intrinsically a very poor conductor and is commonly referred to as being semi-insulating. This property is usually altered by adding dopants of either the p- (positive) or n- (negative) type. This semi-insulating property allows many active devices to be grown on a single substrate, where the semi-insulating GaAs provides the electrical isolation of each device; an important feature in the miniaturization of electronic circuitry, i.e., VLSI (very-large-scale-integration) involving over 100,000 components per chip (one chip is typically between 1 and 10 mm square).

Electron mobility

The higher electron mobility in GaAs than in Si potentially means that in devices where electron transit time is the critical performance parameter, GaAs devices will operate with higher response times than equivalent Si devices. However, the fact that hole mobility is similar for both GaAs and Si means that devices relying on cooperative electron and hole movement, or hole movement alone, show no improvement in response time when GaAs based.

Crystal growth

The bulk crystal growth of GaAs presents a problem of stoichiometric control due the loss, by evaporation, of arsenic both in the melt and the growing crystal ($> ca. 600\text{ }^{\circ}\text{C}$). Melt growth techniques are, therefore, designed to enable an overpressure of arsenic above the melt to be maintained, thus preventing evaporative losses. The loss of arsenic also negates diffusion techniques commonly used for wafer doping in Si technology; since the diffusion temperatures required exceed that of arsenic loss.

Crystal Stress

The thermal gradient and, hence, stress generated in melt grown crystals have limited the maximum diameter of GaAs wafers (currently 6" diameter compared to over 12" for Si), because with increased wafer diameters the thermal stress generated dislocation (crystal imperfections) densities eventually becomes unacceptable for device applications.

Physical strength

Gallium arsenide single crystals are very brittle, requiring that considerably thicker substrates than those employed for Si devices.

Native oxide

Gallium arsenide's native oxide is found to be a mixture of non-stoichiometric gallium and arsenic oxides and elemental arsenic. Thus, the electronic band structure is found to be severely disrupted causing a breakdown in 'normal' semiconductor behavior on the GaAs surface. As a consequence, the GaAs MISFET (metal-insulator-semiconductor-field-effect-transistor) equivalent to the technologically important Si based MOSFET (metal-oxide-semiconductor-field-effect-transistor) is, therefore, presently unavailable.

The passivation of the surface of GaAs is therefore a key issue when endeavoring to utilize the FET technology using GaAs. Passivation in this discussion means the reduction in mid-gap band states which destroy the semiconducting properties of the material. Additionally, this also means the production of a chemically inert coating which prevents the formation of additional reactive states, which can effect the properties of the device.

Bibliography

- S. K. Ghandhi, *VLSI Fabrication Principles: Silicon and Gallium Arsenide*. Wiley-Interscience, New York, (1994).
- *Properties of Gallium Arsenide*. Ed. M. R. Brozel and G. E. Stillman. 3rd Ed. Institution of Electrical Engineers, London (1996).

Synthesis and Purification of Bulk Semiconductors

Introduction

The synthesis and purification of bulk polycrystalline semiconductor material represents the first step towards the commercial fabrication of an electronic device. This polycrystalline material is then used as the raw material for the formation of single crystal material that is processed to semiconductor wafers. The strong influence on the electric characteristics of a semiconductors exhibited by small amounts of some impurities requires that the bulk raw material be of very high purity ($> 99.9999\%$). Although some level of purification is possible during the crystallization process it is important to use as high a purity starting material as possible. While a wide range of substrate materials are available from commercial vendors, silicon and GaAs represent the only large-scale commercial semiconductor substrates, and thus the discussion will be limited to the synthesis and purification of these materials.

Silicon

Following oxygen (46%), silicon (L. silicis flint) is the most abundant element in the earth's crust (28%). However, silicon does not occur in its elemental form, but as its oxide (SiO_2) or as silicates. Sand, quartz, amethyst, agate, flint, and opal are some of the forms in which the oxide appears. Granite, hornblende, asbestos, feldspar, clay and mica, etc. are a few of the numerous silicate minerals. With such boundless supplies of the raw material, the costs associated with the production of bulk silicon is not one of abstraction and conversion of the oxide(s), but of purification of the crude elemental silicon. While 98% elemental silicon, known as metallurgical-grade silicon (MGS), is readily produced on a large scale, the requirements of extreme purity for electronic device fabrication require additional purification steps in order to produce electronic-grade silicon (EGS). Electronic-grade silicon is also known as semiconductor-grade silicon (SGS). In order for the purity levels to be acceptable for subsequent crystal growth and device fabrication, EGS must have carbon and oxygen impurity levels less than a few parts per million (ppm), and metal impurities at the parts per billion (ppb) range or lower. [\[link\]](#) and [\[link\]](#) give typical

impurity concentrations in MGS and EGS, respectively. Besides the purity, the production cost and the specifications must meet the industry desires.

Element	Concentration (ppm)	Element	Concentration (ppm)
aluminum	1000-4350	manganese	50-120
boron	40-60	molybdenum	< 20
calcium	245-500	nickel	10-105
chromium	50-200	phosphorus	20-50
copper	15-45	titanium	140-300
iron	1550-6500	vanadium	50-250
magnesium	10-50	zirconium	20

Typical impurity concentrations found in metallurgical-grade silicon (MGS).

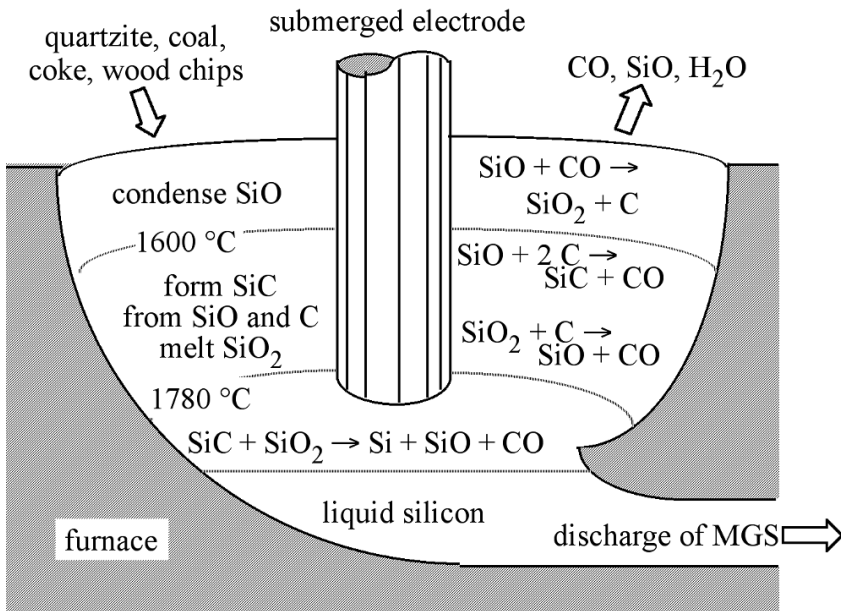
Element	Concentration (ppb)	Element	Concentration (ppb)
arsenic	< 0.001	gold	< 0.00001
antimony	< 0.001	iron	0.1-1.0

boron	≤ 0.1	nickel	0.1-0.5
carbon	100-1000	oxygen	100-400
chromium	< 0.01	phosphorus	≤ 0.3
cobalt	0.001	silver	0.001
copper	0.1	zinc	< 0.1

Typical impurity concentrations found in electronic-grade silicon (EGS).

Metallurgical-grade silicon (MGS)

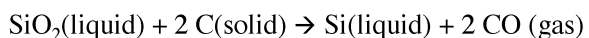
The typical source material for commercial production of elemental silicon is quartzite gravel; a relatively pure form of sand (SiO_2). The first step in the synthesis of silicon is the melting and reduction of the silica in a submerged-electrode arc furnace. An example of which is shown schematically in [\[link\]](#), along with the appropriate chemical reactions. A mixture of quartzite gravel and carbon are heated to high temperatures (ca. 1800 °C) in the furnace. The carbon bed consists of a mixture of coal, coke, and wood chips. The latter providing the necessary porosity such that the gases created during the reaction (SiO and CO) are able to flow through the bed.



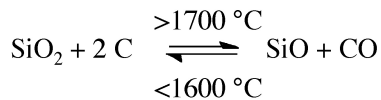
Schematic of submerged-electrode arc furnace for the production of metallurgical-grade silicon (MGS).

The overall reduction reaction of SiO₂ is expressed in [\[link\]](#), however, the reaction sequence is more complex than this overall reaction implies, and involves the formation of SiC and SiO intermediates. The initial reaction between molten SiO₂ and C ([\[link\]](#)) takes place in the arc between adjacent electrodes, where the local temperature can exceed 2000 °C. The SiO and CO thus generated flow to cooler zones in the furnace where SiC is formed ([\[link\]](#)), or higher in the bed where they reform SiO₂ and C ([\[link\]](#)). The SiC reacts with molten SiO₂ ([\[link\]](#)) producing the desired silicon along with SiO and CO. The molten silicon formed is drawn-off from the furnace and solidified.

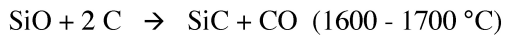
Equation:



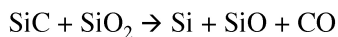
Equation:



Equation:



Equation:



The as-produced MGS is approximately 98-99% pure, with the major impurities being aluminum and iron ([\[link\]](#)), however, obtaining low levels of boron impurities is of particular importance, because it is difficult to remove and serves as a dopant for silicon. The drawbacks of the above process are that it is energy and raw material intensive. It is estimated that the production of one metric ton (1,000 kg) of MGS requires 2500-2700 kg quartzite, 600 kg charcoal, 600-700 kg coal or coke, 300-500 kg wood chips, and 500,000 kWh of electric power. Currently, approximately 500,000 metric tons of MGS are produced per year, worldwide. Most of the production (ca. 70%) is used for metallurgical applications (e.g., aluminum-silicon alloys are commonly used for automotive engine blocks) from whence its name is derived. Applications in a variety of chemical products such as silicone resins account for about 30%, and only 1% or less of the total production of MGS is used in the manufacturing of high-purity EGS for the electronics industry. The current worldwide consumption of EGS is approximately 5×10^6 kg per year.

Electronic-grade silicon (EGS)

Electronic-grade silicon (EGS) is a polycrystalline material of exceptionally high purity and is the raw material for the growth of single-crystal silicon. EGS is one of the purest materials commonly available, see [\[link\]](#). The formation of EGS from MGS is accomplished through chemical purification

processes. The basic concept of which involves the conversion of MGS to a volatile silicon compound, which is purified by distillation, and subsequently decomposed to re-form elemental silicon of higher purity (i.e., EGS). Irrespective of the purification route employed, the first step is physical pulverization of MGS followed by its conversion to the volatile silicon compounds.

A number of compounds, such as monosilane (SiH_4), dichlorosilane (SiH_2Cl_2), trichlorosilane (SiHCl_3), and silicon tetrachloride (SiCl_4), have been considered as chemical intermediates. Among these, SiHCl_3 has been used predominantly as the intermediate compound for subsequent EGS formation, although SiH_4 is used to a lesser extent. Silicon tetrachloride and its lower chlorinated derivatives are used for the chemical vapor deposition (CVD) growth of Si and SiO_2 . The boiling points of silane and its chlorinated products ([\[link\]](#)) are such that they are conveniently separated from each other by fractional distillation.

Compound	Boiling point ($^{\circ}\text{C}$)
SiH_4	-112.3
SiH_3Cl	-30.4
SiH_2Cl_2	8.3
SiHCl_3	31.5
SiCl_4	57.6

Boiling points of silane and chlorosilanes at 760 mmHg (1 atmosphere).

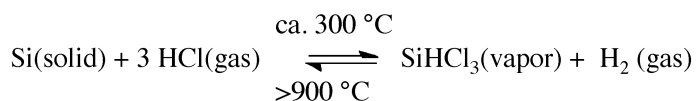
The reasons for the predominant use of SiHCl_3 in the synthesis of EGS are as follows:

1. SiHCl_3 can be easily formed by the reaction of anhydrous hydrogen chloride with MGS at reasonably low temperatures (200 - 400 °C);
2. it is liquid at room temperature so that purification can be accomplished using standard distillation techniques;
3. it is easily handled and if dry can be stored in carbon steel tanks;
4. its liquid is easily vaporized and, when mixed with hydrogen it can be transported in steel lines without corrosion;
5. it can be reduced at atmospheric pressure in the presence of hydrogen;
6. its deposition can take place on heated silicon, thus eliminating contact with any foreign surfaces that may contaminate the resulting silicon;
and
7. it reacts at lower temperatures (1000 - 1200 °C) and at faster rates than does SiCl_4 .

Chlorosilane (Seimens) process

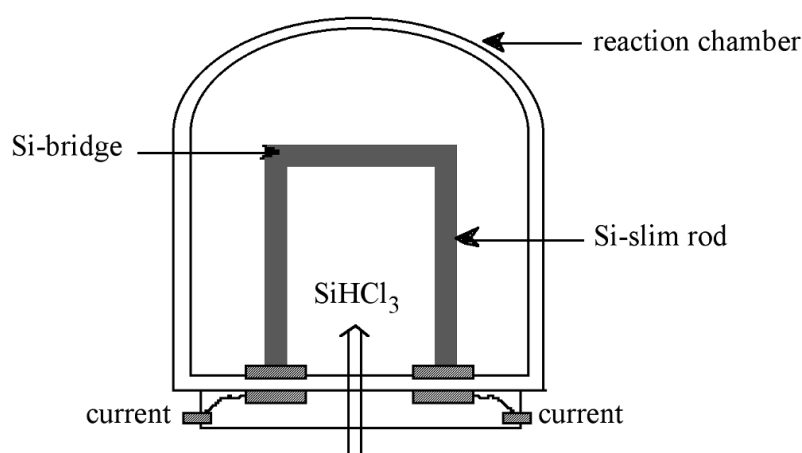
Trichlorosilane is synthesized by heating powdered MGS with anhydrous hydrogen chloride (HCl) at around 300 °C in a fluidized-bed reactor, [\[link\]](#).

Equation:



Since the reaction is actually an equilibrium and the formation of SiHCl_3 highly exothermic, efficient removal of generated heat is essential to assure a maximum yield of SiHCl_3 . While the stoichiometric reaction is that shown in Eq. 5, a mixture of chlorinated silanes is actually prepared which must be separated by fractional distillation, along with the chlorides of any impurities. In particular iron, aluminum, and boron are removed as FeCl_3 (b.p. = 316 °C), AlCl_3 (m.p. = 190 °C subl.), and BCl_3 (b.p. = 12.65 °C), respectively. Fractional distillation of SiHCl_3 from these impurity halides result in greatly increased purity with a concentration of electrically active impurities of less than 1 ppb.

EGS is prepared from purified SiHCl_3 in a chemical vapor deposition (CVD) process similar to the epitaxial growth of Si. The high-purity SiHCl_3 is vaporized, diluted with high-purity hydrogen, and introduced into the Seimens deposition reactor, shown schematically in [\[link\]](#). Within the reactor, thin silicon rods called slim rods (ca. 4 mm diameter) are supported by graphite electrodes. Resistance heating of the slim rods causes the decomposition of the SiHCl_3 to yield silicon, as described by the reverse reaction shown in Eq. 5.



Schematic representation of a Seimens deposition reactor.

The shift in the equilibrium from forming SiHCl_3 from Si at low temperature, to forming Si from SiHCl_3 at high temperature is as a consequence of the temperature dependence ([\[link\]](#)) of the equilibrium constant ([\[link\]](#), where p = partial pressure) for [\[link\]](#). Since the formation of SiHCl_3 is exothermic, i.e., $\Delta H < 0$, an increase in the temperature causes the partial pressure of SiHCl_3 to decrease. Thus, the Siemens process is typically run at ca. 1100 °C, while the reverse fluidized bed process is carried out at 300 °C.

Equation:

$$\ln K_p = \frac{-\Delta H}{RT}$$

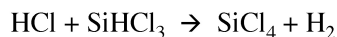
Equation:

$$K_p = \frac{p_{\text{SiHCl}_3} p_{\text{H}_2}}{p_{\text{HCl}}}$$

The slim rods act as a nucleation point for the deposition of silicon, and the resulting polycrystalline rod consists of columnar grains of silicon (polysilicon) grown perpendicular to the rod axis. Growth occurs at less than 1 mm per hour, and after deposition for 200 to 300 hours high-purity (EGS) polysilicon rods of 150-200 mm in diameter are produced. For subsequent float-zone refining the polysilicon EGS rods are cut into long cylindrical rods. Alternatively, the as-formed polysilicon rods are broken into chunks for single crystal growth processes, for example Czochralski melt growth.

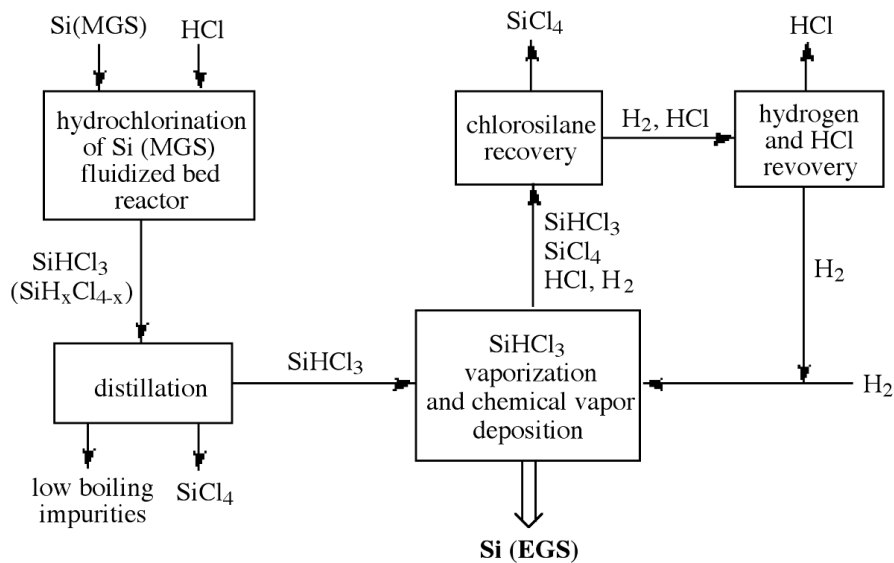
In addition to the formation of silicon, the HCl coproduct reacts with the SiHCl₃ reactant to form silicon tetrachloride (SiCl₄) and hydrogen as major byproducts of the process, [\[link\]](#). This reaction represents a major disadvantage with the Seimens process: poor efficiency of silicon and chlorine consumption. Typically, only 30% of the silicon introduced into CVD reactor is converted into high-purity polysilicon.

Equation:



In order to improve efficiency the HCl, SiCl₄, H₂, and unreacted SiHCl₃ are separated and recovered for recycling. [\[link\]](#) illustrates the entire chlorosilane process starting with MGS and including the recycling of the reaction byproducts to achieve high overall process efficiency. As a consequence, the production cost of high-purity EGS depends on the commercial usefulness of the byproduct, SiCl₄. Additional disadvantages of the Seimens process are derived from its relatively small batch size, slow

growth rate, and high power consumption. These issues have lead to the investigation of alternative cost efficient routes to EGS.

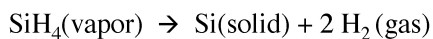


Schematic representation of the reaction pathways for the formation of EGS using the chlorosilane process.

Silane process

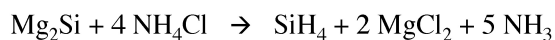
An alternative process for the production of EGS that has begun to receive commercial attention is the pyrolysis of silane (SiH₄). The advantages of producing EGS from SiH₄ instead of SiHCl₃ are potentially lower costs associated with lower reaction temperatures, and less harmful byproducts. Silane decomposes < 900 °C to give silicon and hydrogen, [\[link\]](#).

Equation:



Silane may be prepared by a number of routes, each having advantages with respect to purity and production cost. The simplest process involves the direct reaction of MGS powders with magnesium at 500 °C in a hydrogen atmosphere, to form magnesium silicide (Mg₂Si). The magnesium silicide is then reacted with ammonium chloride in liquid ammonia below 0 °C, [\[link\]](#).

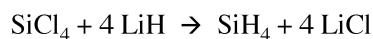
Equation:



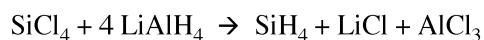
This process is ideally suited to the removal of boron impurities (a p-type dopant in Si), because the diborane (B₂H₆) produced during the reaction forms the Lewis acid-base complex, H₃B(NH₃), whose volatility is sufficiently lower than SiH₄, allowing for the purification of the latter. It is possible to prepare EGS with a boron content of ≤ 20 ppt using SiH₄ synthesized in this manner. However, phosphorus (another dopant) in the form of PH₃ may be present as a contaminant requiring subsequent purification of the SiH₄.

Alternative routes to SiH₄ involve the chemical reduction of SiCl₄ by either lithium hydride ([\[link\]](#)), lithium aluminum hydride ([\[link\]](#)), or via hydrogenation in the presence of elemental silicon ([\[link\]](#) - [\[link\]](#)). The hydride reduction reactions may be carried-out on relatively large scales (ca. 50 kg), but only batch processes. In contrast, Union Carbide has adapted the hydrogenation to a continuous process, involving disproportionation reactions of chlorosilanes ([\[link\]](#) - [\[link\]](#)) and the fractional distillation of silane ([\[link\]](#)).

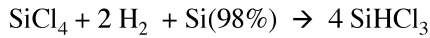
Equation:



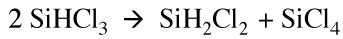
Equation:



Equation:



Equation:



Equation:



Equation:



Pyrolysis of silane on resistively heated polysilicon filaments at 700-800 °C yields polycrystalline EGS. As noted above, the EGS formed has remarkably low boron impurities compared with material prepared from trichlorosilane. Moreover, the resulting EGS is less contaminated with transition metals from the reactor container because SiH_4 decomposition does not cause as much of a corrosion problem as is observed with halide precursor compounds.

Granular polysilicon deposition

Both the chlorosilane (Seimens) and silane processes result in the formation of rods of EGS. However, there has been increased interest in the formation of granular polycrystalline EGS. This process was developed in 1980's, and relies on the decomposition of SiH_4 in a fluidized-bed deposition reactor to produce free-flowing granular polysilicon.

Tiny silicon particles are fluidized in a SiH_4/H_2 flow, and act as seed crystal onto which polysilicon deposits to form free-flowing spherical particles. The size distribution of the particles thus formed is over the range from 0.1 to 1.5 mm in diameter with an average particle size of 0.7 mm. The fluidized-bed

seed particles are originally made by grinding EGS in a ball (or hammer) mill and leaching the product with acid, hydrogen peroxide, and water. This process is time-consuming and costly, and tended to introduce undesirable impurities from the metal grinders. In a new method, large EGS particles are fired at each other by a high-speed stream of inert gas and the collision breaks them down into particles of suitable size for a fluidized bed. This process has the main advantage that it introduces no foreign materials and requires no leaching or other post purification.

The fluidized-bed reactors are much more efficient than traditional rod reactors as a consequence of the greater surface area available during CVD growth of silicon. It has been suggested that fluidized-bed reactors require $1/5$ to $1/10$ the energy, and half the capital cost of the traditional process. The quality of fluidized-bed polysilicon has proven to be equivalent to polysilicon produced by the conventional methods. Moreover, granular EGS in a free-flowing form, and with high bulk density, enables crystal growers to obtain the high, reproducible production yields out of each crystal growth run. For example, in the Czochralski crystal growth process, crucibles can be quickly and easily filled to uniform loading with granular EGS, which typically exceed those of randomly stacked polysilicon chunks produced by the Siemens silane process.

Zone refining

The technique of zone refining is used to purify solid materials and is commonly employed in metallurgical refining. In the case of silicon may be used to obtain the desired ultimate purity of EGS, which has already been purified by chemical processes. Zone refining was invented by Pfann, and makes use of the fact that the equilibrium solubility of any impurity (e.g., Al) is different in the solid and liquid phases of a material (e.g., Si). For the dilute solutions, as is observed in EGS silicon, an equilibrium segregation coefficient (k_0) is defined by $k_0 = C_s/C_l$, where C_s and C_l are the equilibrium concentrations of the impurity in the solid and liquid near the interface, respectively.

If k_0 is less than 1 then the impurities are left in the melt as the molten zone is moved along the material. In a practical sense a molten zone is established in a solid rod. The zone is then moved along the rod from left to right. If $k < 1$ then the frozen part left on the trailing edge of the moving molten zone will be purer than the material that melts in on the right-side leading edge of the moving molten zone. Consequently the solid to the left of the molten zone is purer than the solid on the right. At the completion of the first pass the impurities become concentrated to the right of the solid sample. Repetition of the process allows for purification to exceptionally high levels. [\[link\]](#). lists the equilibrium segregation coefficients for common impurity and dopant elements in silicon; it should be noted that they are all less than 1.

Element	k_0	Element	k_0
aluminum	0.002	iron	8×10^{-6}
boron	0.8	oxygen	0.25
carbon	0.07	phosphorus	0.35
copper	4×10^{-6}	antimony	0.023

Segregation coefficients for common impurity and dopant elements in silicon.

Gallium arsenide

In contrast to electronic grade silicon (EGS), whose use is a minor fraction of the global production of elemental silicon, gallium arsenide (GaAs) is produced exclusively for use in the semiconductor industry. However, arsenic and its compounds have significant commercial applications. The

main use of elemental arsenic is in alloys of Pb, and to a lesser extent Cu, while arsenic compounds are widely used in pesticides and wood preservatives and the production of bottle glass. Thus, the electronics industry represents a minor user of arsenic. In contrast, although gallium has minor uses as a high-temperature liquid seal, manometric fluids and heat transfer media, and for low temperature solders, its main use is in semiconductor technology.

Isolation and purification of gallium metal

At 19 ppm gallium (L. Gallia, France) is about as abundant as nitrogen, lithium and lead; it is twice as abundant as boron (9 ppm), but is more difficult to extract due to the lack of any major gallium-containing ore. Gallium always occurs in association either with zinc or germanium, its neighbors in the periodic table, or with aluminum in the same group. Thus, the highest concentrations (0.1-1%) are in the rare mineral germanite (a complex sulfide of Zn, Cu, Ge, and As), while concentrations in sphalerite (ZnS), diaspore [$\text{AlO}(\text{OH})$], bauxite, or coal, are a hundred-fold less. Industrially, gallium was originally recovered from the flue dust emitted during sulfide roasting or coal burning (up to 1.5% Ga), however, it is now obtained as side product of vast aluminum industry and in particular from the Bayer process for obtaining alumina from bauxite.

The Bayer process involves dissolution of bauxite, $\text{AlO}_x\text{OH}_{3-2x}$, in aqueous NaOH, separation of insoluble impurities, partial precipitation of the trihydrate, $\text{Al}(\text{OH})_3$, and calcination at 1,200 °C. During processing the alkaline solution is gradually enriched in gallium from an initial weight ratio Ga/Al of about 1/5000 to about 1/300. Electrolysis of these extracts with a Hg cathode results in further concentration, and the solution of sodium gallate thus formed is then electrolyzed with a stainless steel cathode to give Ga metal. Since bauxite contains 0.003-0.01% gallium, complete recovery would yield some 500-1000 tons per annum, however present consumption is only 0.1% of this about 10 tons per annum.

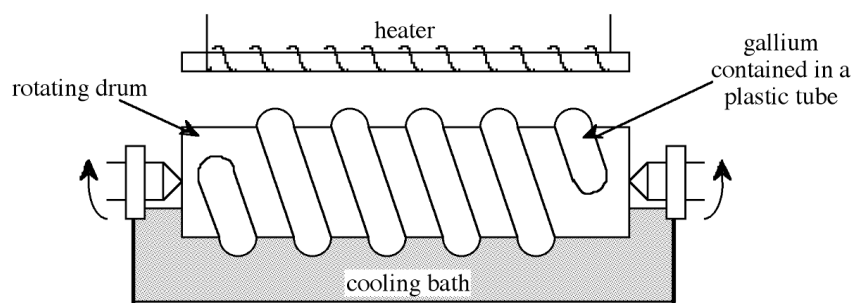
A typical analysis of the 98-99% pure gallium obtained as a side product from the Bayer process is shown in [\[link\]](#). This material is further purified to

99.99% by chemical treatment with acids and O₂ at high temperatures followed by crystallization. This chemical process results in the reduction of the majority of metal impurities at the ppm level, see [\[link\]](#). Purification to seven nines 99.9999% is possible through zone refining, however, since the equilibrium distribution coefficient of the residual impurities $k_0 \approx 1$, multiple passes are required, typically > 500. The low melting point of gallium ensures that contamination from the container wall (which is significant in silicon zone refining) is minimized. In order to facilitate the multiple zone refining in a suitable time, a simple modification of zone refining is employed shown in [\[link\]](#). The gallium is contained in a plastic tube wrapped around a rotating cylinder that is half immersed in a cooling bath. A heater is positioned above the gallium plastic coil. Thus, establishing a series of molten zones that pass upon rotation of the drum by one helical segment per revolution. In this manner, 500 passes may be made in relatively short time periods. The typical impurity levels of gallium zone refined in this manner are given in [\[link\]](#).

Element	Bayer process (ppm)	After acid/base leaching (ppm)	500 zone passes (ppm)
aluminum	100-1,000	7	< 1
calcium	10-100	not detected	not detected
copper	100-1,000	2	< 1
iron	100-1,000	7	< 1
lead	< 2000	30	not detected
magnesium	10-100	1	not detected

mercury	10-100	not detected	not detected
nickel	10-100	not detected	not detected
silicon	10-100	≈ 1	not detected
tin	10-100	≈ 1	not detected
titanium	10-100	1	< 1
zinc	30,000	≈ 1	not detected

Typical analysis of gallium obtained as a side product from the Bayer process.



Schematic representation of a zone refining apparatus.

Isolation and purification of elemental arsenic

Elemental arsenic (L. arsenicum, yellow orpiment) exists in two forms: yellow (cubic, As_4) and gray or metallic (rhombohedral). At a natural abundance of 1.8 ppm arsenic is relatively rare, however, this is offset by its presence in a number of common minerals and the relative ease of isolation.

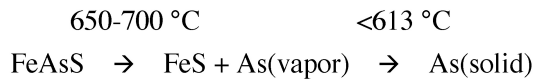
Arsenic containing minerals are grouped into three main classes: the sulfides realgar (As_4S_4) and orpiment (As_2S_3), the oxide arsenolite (As_2O_3), and the arsenides and sulfarsenides of the iron, cobalt, and nickel. Minerals in this latter class include: loellinginite (FeAs_2), safforlite (CoAs), niccolite (NiAs), rammelsbergite (NiAs_2), arsenopyrite or mispickel (FeAsS), cobaltite (CoAsS), enargite (Cu_3AsS_4), gerdorfite (NiAsS), and the quarternary sulfide glaucodot $[(\text{Co,Fe})\text{AsS}]$. [\[link\]](#) shows the typical impurities in arsenopyrite.

Element	Concentration (ppm)	Element	Concentration (ppm)
silver	90	nickel	< 3,000
gold	8	lead	50
cobalt	30,000	platinum	0.4
copper	200	rhenium	50
germanium	30	selenium	50
manganese	3,000	vanadium	300
molybdenum	60	zinc	400

Typical impurities in arsenopyrite.

Arsenic is obtained commercially by smelting either FeAs_2 or FeAsS at 650-700 °C in the absence of air and condensing the sublimed element ($T_{\text{sub}} = 613 \text{ °C}$), [\[link\]](#).

Equation:

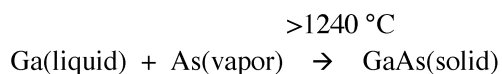


The arsenic thus obtained is combined with lead and then sublimed ($T_{\text{sub}} = 614\text{ }^{\circ}\text{C}$) which binds any sulfur impurities more strongly than arsenic. Any residual arsenic that remains trapped in the iron sulfide is separated by forming the oxide (As_2O_3) by roasting the sulfide in air. The oxide is sublimed into the flue system during roasting from where it is collected and reduced with charcoal at $700-800\text{ }^{\circ}\text{C}$ to give elemental arsenic. Semiconductor grade arsenic ($> 99.9999\%$) is formed by zone refining.

Synthesis and purification of gallium arsenide.

Gallium arsenide can be prepared by the direct reaction of the elements, [\[link\]](#). However, while conceptually simple the synthesis of GaAs is complicated by the different vapor pressures of the reagents and the highly exothermic nature of the reaction. Furthermore, since the synthesis of GaAs at atmospheric pressure is accompanied by its simultaneous decomposes due to the loss by sublimation, of arsenic, the synthesis must be carried out under an overpressure of arsenic in order to maintain a stoichiometric composition of the synthesized GaAs.

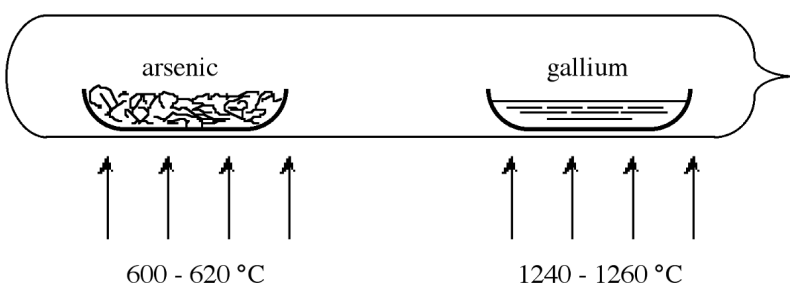
Equation:



In order to overcome the problems associated with arsenic loss, the reaction is usually carried out in a sealed reaction tube. However, if a stoichiometric quantity of arsenic is used in the reaction a constant temperature of $1238\text{ }^{\circ}\text{C}$ must be employed in order to maintain the desired arsenic overpressure of 1 atm. Practically, it is easier to use a large excess of arsenic heated to a lower temperature. In this situation the pressure in the tube is approximately equal to the equilibrium vapor pressure of the volatile component (arsenic) at the lower temperature. Thus, an over pressure of 1 atm arsenic may be

maintained if within a sealed tube elemental arsenic is heated to 600-620 °C while the GaAs is maintained at 1240-1250 °C.

[\[link\]](#) shows the sealed tube configuration that is typically used for the synthesis of GaAs. The tube is heated within a two-zone furnace. The boats holding the reactants are usually made of quartz, however, graphite is also used since the latter has a closer thermal expansion match to the GaAs product. If higher purity is required then pyrolytic boron nitride (PBN) is used. One of the boats is loaded with pure gallium the other with arsenic. A plug of quartz wool may be placed between the boats to act as a diffuser. The tube is then evacuated and sealed. Once brought to the correct reaction temperatures ([\[link\]](#)), the arsenic vapor is transported to the gallium, and they react to form GaAs in a controlled manner. [\[link\]](#) gives the typical impurity concentrations found in polycrystalline GaAs.



Schematic representation of a sealed tube synthesis of GaAs.

Element	Concentration (ppm)	Element	Concentration (ppm)
boron	0.1	silicon	0.02

carbon	0.7	phosphorus	0.1
nitrogen	0.1	sulfur	0.01
oxygen	0.5	chlorine	0.08
fluorine	0.2	nickel	0.04
magnesium	0.02	copper	0.01
aluminum	0.02	zinc	0.05

Impurity concentrations found in polycrystalline GaAs.

Polycrystalline GaAs, formed in from the direct reaction of the elements is often used as the starting material for single crystal growth via Bridgeman or Czochralski crystal growth. It is also possible to prepare single crystals of GaAs directly from the elements using in-situ, or direct, compounding within a high-pressure liquid encapsulated Czochralski (HPLEC) technique.

Bibliography

- K. G. Baraclough, K. G., in *The Chemistry of the Semiconductor Industry*, Eds. S. J. Moss and A. Ledwith, Blackie and Sons, Glasgow, Scotland (1987).
- L. D. Crossman and J. A. Baker, *Semiconductor Silicon 1977*, Electrochem. Soc., Princeton, New Jersey (1977).
- M. Fleisher, in *Economic Geology, 50th Aniv. Vol.*, The Economic Geology Publishing Company, Lancaster, PA (1955).
- G. Hsu, N. Rohatgi, and J. Houseman, *AIChE J.*, 1987, **33**, 784.
- S. K. Iya, R. N. Flagella, and F. S. Dipaolo, *J. Electrochem. Soc.*, 1982, **129**, 1531.
- J. Krauskopf, J. D. Meyer, B. Wiedemann, M. Waldschmidt, K. Bethge, G. Wolf, and W. Schültze, 5th Conference on Semi-insulating III-V Materials, Malmo, Sweden, 1988, Eds. G. Grossman and L. Ledebo, Adam-Hilger, New York (1988).

- J. R. McCormic, Conf. Rec. 14th IEEE Photovolt. Specialists Conf., San Diego, CA (1980).
- J. R. McCormic, in *Semiconductor Silicon 1981*, Ed. H. R. Huff, Electrochemical Society, Princeton, New Jersey (1981).
- W. C. O'Mara, Ed. *Handbook of Semiconductor Silicon Technology*, Noyes Pub., New Jersey (1990).
- W. G. Pfann, *Zone Melting*, John Wiley & Sons, New York, (1966).
- F. Shimura, *Semiconductor Silicon Crystal Technology*, Academic Press (1989).

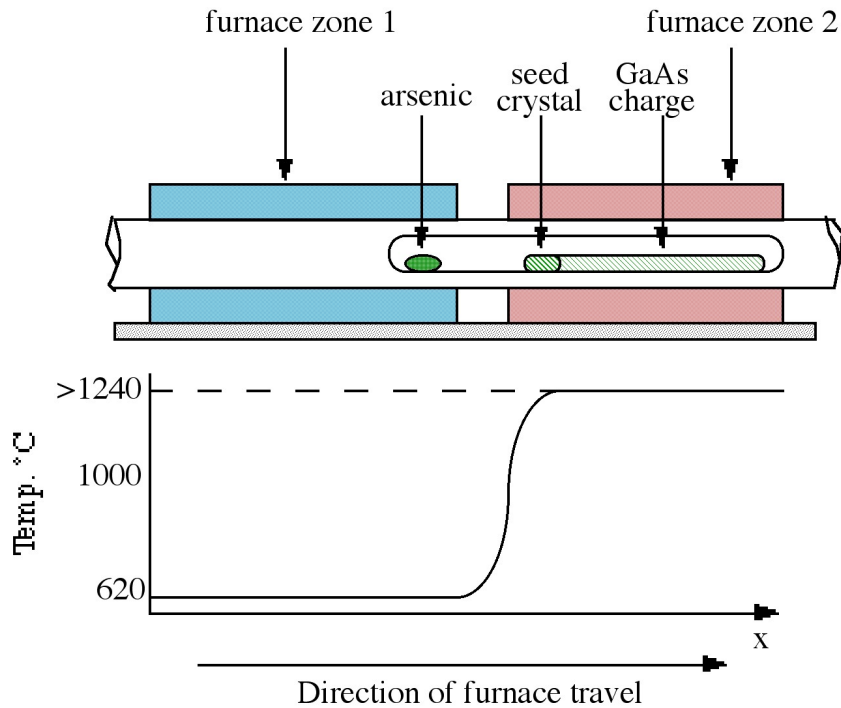
Growth of Gallium Arsenide Crystals

Introduction

When considering the synthesis of Group 13-15 compounds for electronic applications, the very nature of semiconductor behavior demands the use of high purity single crystal materials. The polycrystalline materials synthesized above are, therefore, of little use for 13-15 semiconductors but may, however, serve as the starting material for melt grown single crystals. For GaAs, undoubtedly the most important 13-15 (III - V) semiconductor, melt grown single crystals are achieved by one of two techniques: the Bridgman technique, and the Czochralski technique.

Bridgman growth

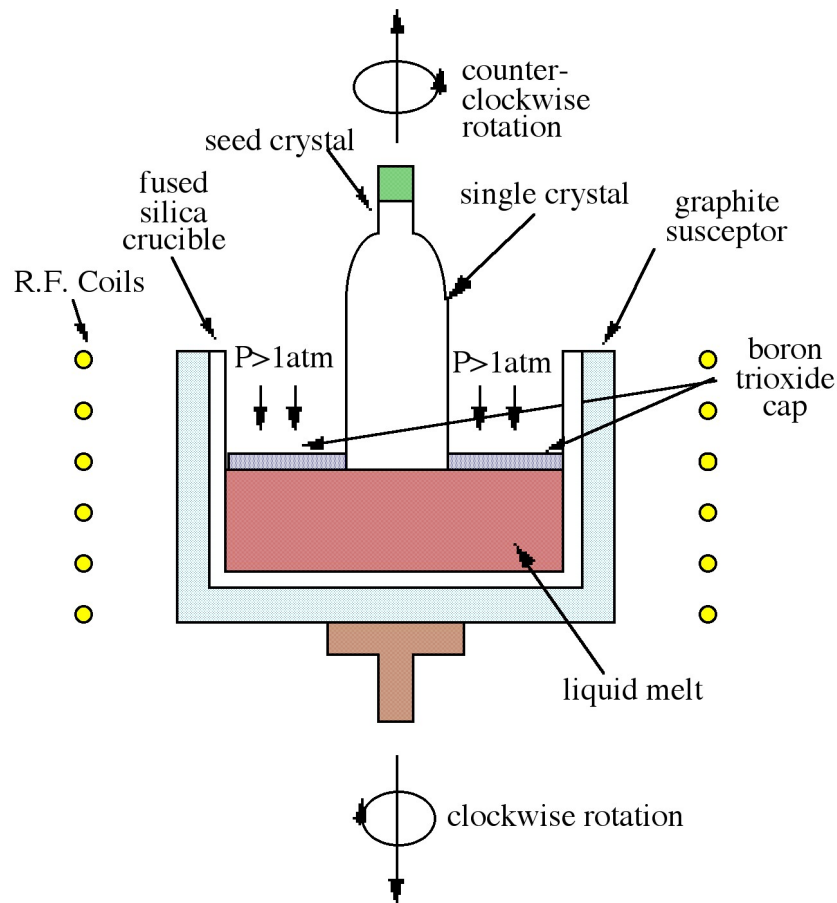
The Bridgman technique requires a two-zone furnace, of the type shown in [\[link\]](#). The left hand zone is maintained at a temperature of *ca.* 610 °C, allowing sufficient overpressure of arsenic within the sealed system to prevent arsenic loss from the gallium arsenide. The right hand side of the furnace contains the polycrystalline GaAs raw material held at a temperature just above its melting point (*ca.* 1240 °C). As the furnace moves from left to right, the melt cools and solidifies. If a seed crystal is placed at the left hand side of the melt (at a point where the temperature gradient is such that only the end melts), a specific orientation of single crystal may be propagated at the liquid-solid interface eventually to produce a single crystal.



A schematic diagram of a Bridgman two-zone furnace used for melt growths of single crystal GaAs.

Czochralski growth

The Czochralski technique, which is the most commonly used technique in industry, is shown in [\[link\]](#). The process relies on the controlled withdrawal of a seed crystal from a liquid melt. As the seed is lowered into the melt, partial melting of the tip occurs creating the liquid solid interface required for crystal growth. As the seed is withdrawn, solidification occurs and the seed orientation is propagated into the grown material. The variable parameters of rate of withdrawal and rotation rate can control crystal diameter and purity. As shown in [\[link\]](#) the GaAs melt is capped by boron trioxide (B_2O_3). The capping layer, which is inert to GaAs, prevents arsenic loss when the pressure on the surface is above atmospheric pressure. The growth of GaAs by this technique is thus termed liquid encapsulated Czochralski (LEC) growth.



A schematic diagram of the Czochralski technique as used for growth of GaAs single crystal bond.

While the Bridgman technique is largely favored for GaAs growth, larger diameter wafers can be obtained by the Czochralski method. Both of these melt techniques produce materials heavily contaminated by the crucible, making them suitable almost exclusively as substrate material. Another disadvantage of these techniques is the production of defects in the material caused by the melt process.

Bibliography

- W. G. Pfann, *Zone Melting*, John Wiley & Sons, New York (1966).

- R. E. Williams, *Gallium Arsenide Processing Techniques*. Artech House (1984).

Ceramic Processing of Alumina

Introduction

While aluminum is the most abundant metal in the earth's crust (*ca.* 8%) and aluminum compounds such as alum, $K[Al(SO_4)_2] \cdot 12(H_2O)$, were known throughout the world in ancient times, it was not until the isolation of aluminum in the late eighteenth century by the Danish scientist H. C. Öersted that research into the chemistry of the Group 13 elements began in earnest. Initially, metallic aluminum was isolated by the reduction of aluminum trichloride with potassium or sodium; however, with the advent of inexpensive electric power in the late 1800's, it became economically feasible to extract the metal *via* the electrolysis of alumina (Al_2O_3) dissolved in cryolite, Na_3AlF_6 , (the Hall-Heroult process). Today, alumina is prepared by the Bayer process, in which the mineral bauxite (named for Les Baux, France, where it was first discovered) is dissolved with aqueous hydroxides, and the solution is filtered and treated with CO_2 to precipitate alumina. With availability of both the mineral and cheap electric power being the major considerations in the economical production of aluminum, it is not surprising that the leading producers of aluminum are the United States, Japan, Australia, Canada, and the former Soviet Union.

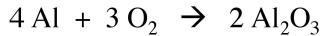
Aluminum oxides and hydroxides

The many forms of aluminum oxides and hydroxides are linked by complex structural relationships. Bauxite has the formula $Al_x(OH)_{3-2x}$ ($0 < x < 1$) and is thus a mixture of Al_2O_3 (α -alumina), $Al(OH)_3$ (gibbsite), and $AlO(OH)$ (boehmite). The latter is an industrially important compound which is used in the form of a gel as a pre-ceramic in the production of fibers and coatings, and as a fire retarding agent in plastics.

Heating boehmite and diaspor to 450 °C causes dehydration to yield forms of alumina which have structures related to their oxide-hydroxide precursors. Thus, boehmite produces the low-temperature form γ -alumina, while heating diaspor will give α -alumina (corundum). γ -alumina converts to the hcp structure at 1100 °C. A third form of Al_2O_3 forms on the surface of the clean aluminum metal. The thin, tough, transparent oxide layer is the

reason for much of the usefulness of aluminum. This oxide skin is rapidly self-repairing because its heat of formation is so large ($\Delta H = -3351 \text{ kJ/mol}$).

Equation:



Ternary and mixed-metal oxides

A further consequence of the stability of alumina is that most if not all of the naturally occurring aluminum compounds are oxides. Indeed, many precious gemstones are actually corundum doped with impurities.

Replacement of aluminum ions with trace amounts of transition-metal ions transforms the formerly colorless mineral into ruby (red, Cr^{3+}), sapphire (blue, $\text{Fe}^{2+/3+}$, Ti^{4+}), or topaz (yellow, Fe^{3+}). The addition of stoichiometric amounts of metal ions causes a shift from the $\alpha\text{-Al}_2\text{O}_3$ hcp structure to the other common oxide structures found in nature. Examples include the perovskite structure for ABO_3 type minerals (e.g., CeTiO_7 or LaAlO_3) and the spinel structure for AB_2O_4 minerals (e.g., beryl, BeAl_2O_4).

Aluminum oxide also forms ternary and mixed-metal oxide phases. Ternary systems such as mullite ($\text{Al}_6\text{Si}_2\text{O}_{13}$), yttrium aluminum garnet (YAG, $\text{Y}_3\text{Al}_5\text{O}_{12}$), the β -aluminas (e.g., $\text{NaAl}_{11}\text{O}_{17}$) and aluminates such as hibonite ($\text{CaAl}_{12}\text{O}_{19}$) possessing β -alumina or magnetoplumbite-type structures can offer advantages over those of the binary aluminum oxides.

Applications of these materials are found in areas such as engineering composite materials, coatings, technical and electronic ceramics, and catalysts. For example, mullite has exceptional high temperature shock resistance and is widely used as an infrared-transparent window for high temperature applications, as a substrate in multilayer electronic device packaging, and in high temperature structural applications. Hibonite and other hexaluminates with similar structures are being evaluated as interfacial coatings for ceramic matrix composites due to their high thermal stability and unique crystallographic structures. Furthermore, aluminum oxides doped with an alkali, alkaline earth, rare earth, or transition metal are

of interest for their enhanced chemical and physical properties in applications utilizing their unique optoelectronic properties.

Synthesis of aluminum oxide ceramics

In common with the majority of oxide ceramics, two primary synthetic processes are employed for the production of aluminum oxide and mixed metal oxide materials:

- 1. The traditional ceramic powder process.
- 2. The solution-gelation, or "sol-gel" process.

The environmental impact of alumina and alumina-based ceramics is in general negligible; however, the same cannot be said for these methods of preparation. As practiced commercially, both of the above processes can have a significant detrimental environmental impact.

Traditional ceramic processing

Traditional ceramic processing involves three basic steps generally referred to as powder-processing, shape-forming, and densification, often with a final mechanical finishing step. Although several steps may be energy intensive, the most direct environmental impact arises from the shape-forming process where various binders, solvents, and other potentially toxic agents are added to form and stabilize a solid ("green") body ([link](#)).

Function	Composition	Volume (%)
Powder	alumina (Al ₂ O ₃)	27

Solvent	1,1,1-trichloroethane/ethanol	58
Deflocculant	menhaden oil	1.8
Binder	poly(vinyl butyrol)	4.4
Plasticizer	poly(ethylene glycol)/octyl phthalate	8.8

Typical composition of alumina green body

The component chemicals are mixed to a slurry, cast, then dried and fired. In addition to any innate health risk associated with the chemical processing these agents are subsequently removed in gaseous form by direct evaporation or pyrolysis. The replacement of chlorinated solvents such as 1,1,1-trichloroethylene (TCE) must be regarded as a high priority for limiting environmental pollution. The United States Environmental Protection Agency (EPA) included TCE on its 1991 list of 17 high-priority toxic chemicals targeted for source reduction. The plasticizers, binders, and alcohols used in the process present a number of potential environmental impacts associated with the release of combustion products during firing of the ceramics, and the need to recycle or discharge alcohols which, in the case of discharge to waterways, may exert high biological oxygen demands in the receiving communities. It would be desirable, therefore, to be able to use aqueous processing; however, this has previously been unsuccessful due to problems associated with batching, milling, and forming. Nevertheless, with a suitable choice of binders, etc., aqueous processing is possible. Unfortunately, in many cast-parts formed by green body processing the liquid solvent alone consists of over 50 % of the initial volume, and while this is not directly of an environmental concern, the resultant shrinkage makes near net shape processing difficult.

Sol-gel

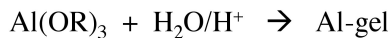
Whereas the traditional sintering process is used primarily for the manufacture of dense parts, the solution-gelation (sol-gel) process has been

applied industrially primarily for the production of porous materials and coatings.

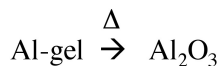
Sol-gel involves a four stage process: dispersion, gelation, drying, and firing. A stable liquid dispersion or *sol* of the colloidal ceramic precursor is initially formed in a solvent with appropriate additives. By changing the concentration (aging) or pH, the dispersion is "polymerized" to form a solid dispersion or *gel*. The excess liquid is removed from this gel by drying and the final ceramic is formed by firing the gel at higher temperatures.

The common sol-gel route to aluminum oxides employs aluminum hydroxide or hydroxide-based material as the solid colloid, the second phase being water and/or an organic solvent, however, the strong interactions of the freshly precipitated alumina gels with ions from the precursor solutions makes it difficult to prepare these gels in pure form. To avoid this complication, alumina gels are also prepared from the hydrolysis of aluminum alkoxides, Al(OR)_3 .

Equation:



Equation:



The exact composition of the gel in commercial systems is ordinarily proprietary, however, a typical composition will include an aluminum compound, a mineral acid, and a complexing agent to inhibit premature precipitation of the gel, e.g., [\[link\]](#).

--	--

Function	Composition
Boehmite precursor	ASB [aluminum <i>sec</i> -butoxide, $\text{Al}(\text{OC}_4\text{H}_9)_3$]
Electrolyte	HNO_3 0.07 mole/mole ASB
Complexing agent	glycerol <i>ca.</i> 10 wt.%

Typical composition of an alumina sol-gel for slipcast ceramics.

The principal environmental consequences arising from the sol-gel process are those associated with the use of strong acids, plasticizers, binders, solvents, and *sec*-butanol formed during the reaction. Depending on the firing conditions, variable amounts of organic materials such as binders and plasticizers may be released as combustion products. NO_x 's may also be produced in the off-gas from residual nitric acid or nitrate salts. Moreover, acids and solvents must be recycled or disposed of. Energy consumption in the process entails "upstream" environmental emissions associated with the production of that energy.

Bibliography

- *Advances in Ceramics*, Eds. J. A. Mangels and G. L. Messing, American Ceramic Society, Westville, OH, 1984, Vol. 9.
- Adkins, *J. Am. Chem. Soc.*, 1922, **44**, 2175.
- A. R. Barron, *Comm. Inorg. Chem.*, 1993, **14**, 123.
- M. K. Cinibulk, *Ceram. Eng. Sci., Proc.*, 1994, **15**, 721.
- F. A. Cotton and G. Wilkinson, *Advanced Inorganic Chemistry*, 5th Ed., John Wiley and Sons, New York (1988).
- N. N. Greenwood and A. Earnshaw, *Chemistry of the Elements*, Pergamon Press, Oxford (1984).
- P. H. Hsu and T. F. Bates, *Mineral Mag.*, 1964, **33**, 749.
- W. D. Kingery, H. K. Bowen, and D. R. Uhlmann, *Introduction to Ceramics*, 2nd Ed. Wiley, New York (1976).
- H. Schneider, K. Okada, and J. Pask, *Mullite and Mullite Ceramics*, Wiley (1994).

- R. V. Thomas, *Systems Analysis and Water Quality Management*, McGraw-Hill, New York (1972).
- J. C. Williams, in *Treatise on Materials Science and Technology*, Ed. F. F. Y. Wang, Academic Press, New York (1976).

Piezoelectric Materials Synthesis

This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Ilse Y. Guzman-Jimenez.

Introduction

Piezoelectricity is the generation of an electric moment by a change of stress applied to a solid. The word piezoelectricity literally means “pressure electricity”; the prefix piezo is derived from the Greek word *piezein*, “to press”. The piezoelectric effect was discovered in 1880 by the brothers Jacques and Pierre Curie. Not only did they demonstrate the phenomenon, but they also established the criteria for its existence in a given crystal. Of the thirty-two crystal classes, twenty-one are non-centrosymmetric (not having a centre of symmetry), and of these, twenty exhibit direct piezoelectricity.

The first practical application of the piezoelectric effect was developed when ground quartz crystals were placed between the plates of a tuning capacitor in order to stabilize oscillating circuits in radio transmitters and receivers; however, the phenomenon of piezoelectricity was not well exploited until World War I, when Langevin used piezoelectrically excited quartz plates to generate sound waves in water for use in submarine detection.

Piezoelectricity can also occur in polycrystalline or amorphous substances which have become anisotropic by external agents. Synthetic piezoelectric materials became available near the end of World War II, with the accidental discovery of the fact that materials like barium titanate and rare earth oxides become piezoelectric when they are polarized electrically. During the postwar years, when germanium and silicon were revolutionizing the electronics industry, piezoceramics appeared for a while to be joining the revolution, but the limited availability of materials and components, made the piezoelectric phenomenon failed to lead mature applications during the 1950s. It is only now that a variety of piezoelectric materials are being synthesized and optimized. As a consequence piezoelectric-based devices are undergoing a revolutionary development, specially for medicine and aerospace applications.

Piezoelectric ceramics

Most piezoelectric transducers are made up of ceramic materials for a broad range of electromechanical conversion tasks as transmitters, ranging from buzzers in alarm clocks to sonars, and as receivers, ranging from ultra high frequency (UHF) filters to hydrophones.

Most of the piezoelectric materials in usage are from the lead zirconate titanate (PZT) family, because of their excellent piezoelectric parameters, thermal stability, and dielectric properties. Additionally the properties of this family can be modified by changing the zirconium to titanium ratio or by addition of both metallic and non-metallic elements. PZT ($\text{PbZr}_{1-x}\text{Ti}_x\text{O}_3$) ceramics and their solid solutions with several complex perovskite oxides have been studied; among the various complex oxide materials, niobates have attracted special attention. Ternary ceramic materials, lead metaniobate, as well as, barium and modified lead titanates complete the list of piezoceramic materials.

Selective parameters for piezoceramic materials are given in [\[link\]](#), where Q_m is the mechanical quality factor, T_c is the Curie point, d_{31} is the the transverse charge coefficient, and k_p , k_t and k_{31} are the electromechanical coupling factors for planar, thickness, and transversal mode respectively.

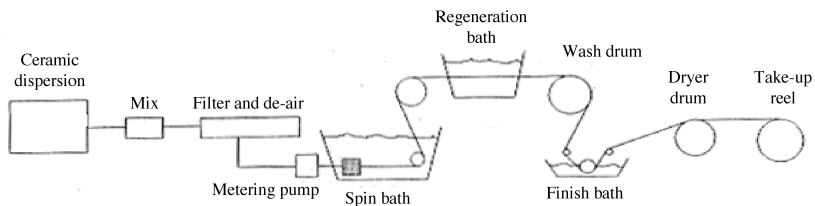


Material property	PZT modified	Lead metaniobate	PSZNT 31/40/29	PZT, x = 0.5	PSN-PLT	TsTS-42-1 50/50	PZT, x = 0.48
Q_m	350	40	222	74	41	887	
T_c (°C)	290	462		369	152		355
$d_{31}(\times 10^{-12} \text{ C/N})$					-79	50	
k_p	0.5		60	0.428	30.7	46.5	
k_t		0.32		0.438	-		
k_{31}		0.21		0.263	17.9		

Selective parameters for illustrative piezoceramic materials.

Recently, sol-gel processing has been used to prepare ceramics, making possible the preparation of materials that are difficult to obtain by conventional methods. Both, inorganic and organic precursor have been reported. Additionally, new techniques for the production of ceramic fibers have been developed. Better processing and geometrical and microstructural control are the main goals in the production of fibers.

The latest development in piezoceramic fibers is the modification of the viscous-suspension-spinning process (VSSP) for the production of continuous piezoelectric ceramic fibers for smart materials and active control devices, such as transducers, sensor/actuators and structural-control devices. The VSSP utilizes conventional synthesized ceramic powders and cellulose, as the fugitive carrier, to produce green ceramic fiber at a reasonable cost. [\[link\]](#) shows the schematic representation of the VSSP.



The viscous-suspension-spinning process (VSSP) for the production of continuous piezoceramic fiber.

Synthesis of reactive PZT precursor powder by the oxalate coprecipitation technique has also been developed. The precursor transforms to phase pure PZT at or above 850 °C the PZT obtained by this technique showed a Curie temperature of 355 °C. The advantages of the coprecipitation technique are the lack of moisture sensitive and special handling precursors.

Although new materials have been investigated with the purpose of create replacements for ceramics, there has been a great improvement in their properties and, current research is focused in the development of

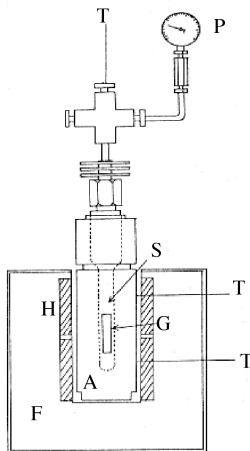
new techniques for both synthesis and processing.

Piezoelectric single crystals.

The recent progress of the electronic technology requires new piezoelectric crystals with a high thermal stability and large electromechanical coupling factors. Single-crystal materials have been considered as replacements for polycrystalline ceramics. Ideally single-crystals of lead zirconate titanate (PZT) itself would be the main choice as it is the most prevailing piezoelectric material, but it is difficult to grow large single crystals. On the other hand, the fact that single-crystals offer many advantages over polycrystalline systems has been recognized. Materials such as lithium niobate present essentially no aging, no mechanical creep and excellent performance in high temperature conditions.

New piezoelectric single crystals grown by conventional RF-heating Czochralski (CZ) technique have been synthesized. High purity starting materials, mainly oxides powders, and Ar atmosphere are required. $\text{La}_3\text{Ga}_5\text{SiO}_{14}$, $\text{La}_3\text{Nb}_{0.5}\text{Ga}_{5.5}\text{O}_{14}$ and $\text{La}_3\text{Ta}_{0.5}\text{Ga}_{5.5}\text{O}_{14}$ single crystals have been grown by using this method. However, the CZ technique can be applied only to materials that can be synthesized by ordinary solid-state reaction and can undergo the pulling method.

$\text{BaBe}_2\text{Si}_2\text{O}_7$ (barylite) has been known as material with a strong piezoelectricity, however, it can not be obtained by solid-state reaction and CZ technique therefore is not applicable. As an alternative for piezoelectric crystals growth hydrothermal synthesis has been developed. [\[link\]](#) shows the experimental apparatus for the growth of barylite. Eventhough, crystals can be obtained using this technique, high pressure (500 - 1000 bar) and a solvent for the raw materials are required.



Experimental apparatus for the hydrothermal synthesis of barylite. H = heater, F = furnace, S = specimen vessel, G = growth capsule, P = pressure gauge, and T = thermocouples.

Adapted from M.
Maeda, T. Uehara,
H. Sato and T.
Ikeda, *Jpn. J. Appl.
Phys.*, 1991, **30**,
2240.

While the piezoceramics dominate the single crystal materials in usage, single crystals piezoelectrics continue to make important contributions both in price-conscious consumer market and in performance-driven defense applications. Areas such as frequency stabilized oscillators, surface acoustic wave devices and filters with a wide pass band, are still dominated by single crystals.

Piezoelectric thin films

Recently, there has been great interest in the deposition of piezoelectric thin films, mainly for microelectronic systems (MEMS) applications; where the goal is to integrate sensors and actuators based on PZT films with Si semiconductor-based signal processing; and for surface acoustic wave (SAW) devices; where the goal is to achieve higher electromechanical coupling coefficient and temperature stability. Piezoelectrical microcantilevers, microactuators, resonators and SAW devices using thin films have been reported.

Several methods have been investigated for PZT thin films. In the metallo-organic thin film deposition, alkoxides are stirred during long periods of time (up to 18 hours). After pyrolysis, PZT amorphous films are formed and then calcination between 400 – 600 °C for 80 hours leads to PZT crystallization (perovskite phase) by a consecutive phase transformation process, which involves a transitional pyrochlore phase.

A hybrid metallorganic decomposition (MOD) route has also been developed to prepare PZT thin films. Lead and titanium acetates and, zirconium acetylacetonate are used. The ferroelectric piezoelectric and dielectric properties indicate that the MOD route provides PZT films of good quality and comparable to literature values. In addition to being simple, MOD has several advantages which include: homogeneity at molecular level and ease composition control.

Metalorganic chemical vapor deposition (MOCVD) has been applied to PZT thin films deposition also. It has been proved that excellent quality PZT films can be grown by using MOCVD, but just recently the control of microstructure the deposition by varying the temperature, Zr to Ti ratio and precursors flow has been studied. Recent progress in PZT films deposition has led to lower temperature growth and it is expected that by lowering the deposition temperature better electrical properties can be achieved. Additionally, novel techniques such as KrF excimer laser ablation and, ion and photo-assisted depositions, have also been used for PZT films synthesis.

On the other hand, a single process to deposit PZT thin film by a hydrothermal method has been reported recently. Since the sol-gel method, sputtering and chemical vapor deposition techniques are useful only for making flat materials, the hydrothermal method offers the advantage of making curved shaped materials. The hydrothermal method utilizes the chemical reaction between titanium and ions melted in solution. A PZT thin film has been successfully deposited directly on a titanium substrate and the optimum ion ratio in the solution is being investigated to improve the piezoelectric effect.

Among the current reported piezoelectric materials, the $\text{Pb}(\text{Ni}_{1/3}\text{Nb}_{2/3})_{0.2}\text{Zr}_{0.4}\text{Ti}_{0.4}\text{O}_3$ (PNNZT, 2/4/4) ferroelectric ceramic has piezoelectric properties that are about 60 and 3 times larger than the reported values for ZnO and PZT. A sol-gel technique has been developed for the deposition of a novel piezoelectric

PNNZT thin film. A 2-methoxyethanol based process is used. In this process precursors are heated at lower temperature than the boiling point of the solvent, to distill off water. Then prior high temperature annealing, addition of excess Pb precursor in the precursor solution is required to compensate the lead loss. The pure perovskite phase is then obtained at 600 °C, after annealing.

Thin films of zinc oxide (ZnO), a piezoelectric material and n-type wide-bandgap semiconductor, have been deposited. ZnO films are currently used in SAW devices and in electro-optic modulators. ZnO thin films have been grown by chemical vapor deposition and both d.c. and r.f. sputtering techniques. Recently, optimization of ZnO films by r.f. magnetron sputtering has been developed. However, homogeneity is one of the main problems when using this technique, since films grown by this optimized method, showed two regions with different piezoelectric properties.

DC magnetron sputtering is other technique for piezoelectric thin film growth, recently aluminum nitride, a promising material for use in thin-film bulk acoustic wave resonators for applications in RF bandpass filters, has been grown by this method. The best quality films are obtained on Si substrates. In order to achieve the highest resonator coupling, the AlN must be grown directly on the electrodes. The main problem in the AlN growth is the oxygen contamination, which leads to the formation of native oxide on the Al surface, preventing crystalline growth of AlN.

Piezoelectric polymers

The discovery of piezoelectricity in polymeric materials such as polyvinylidene difluoride (PVF), was considered as an indication of a renaissance in piezoelectricity. Intensive research was focused in the synthesis and functionalization of polymers. A potential piezoelectric polymer has to contain a high concentration of dipoles and also be mechanically strong and film-forming. The degree of crystallinity and the morphology of the crystalline material have profound effects on the mechanical behavior of polymers. Additionally, in order to induce a piezoelectric response in amorphous systems the polymer is poled by application of a strong electric field at elevated temperature sufficient to allow mobility of the molecular dipoles in the polymer. Recent approaches have been focused in the development of cyano-containing polymers, due to the fact that cyano polymers could have many dipoles which can be aligned in the same direction.

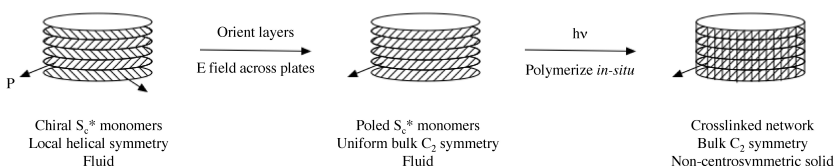
Phase transfer catalyzed reaction has been used for piezoelectric polymer preparation from malonitrile, however this method leads to low molecular weight, and low yield of impure vinylidene cyanide units containing material. The use of solid K_2CO_3 and acetonitrile without added phase transfer catalyst shows excellent yields for polyester possessing backbone gem-dinitriles and for polyamide synthesis. The polyester and polyamide obtained contained a dinitrile group net dipole which can be align in the same direction as the carbonyl groups.

The pursuit for better piezoelectric polymers has led to molecular modeling which indicates that one cyano substituent should be almost as effective as two geminal cyano substituents, opening a new area of potential materials having an acrylonitrile group as the basic building block. However, polyacrylonitrile itself is not suitable because it forms a helix. Thus acrylonitrile copolymers have been investigated.

Most of the piezoelectric polymers available are still synthesized by conventional methods such as polycondensation and radical polymerization. Therefore piezoelectric polymer synthesis has the same problems as the commercial polymer preparation, such as controlling the degree of polymerization and crystallinity.

A novel technique of vapor deposition polymerization has been reported as an alternative method to copolymeric thin films. Aliphatic polyurea 9 was synthesized by evaporating monomers of 1,9-diaminononano and 1,9-diisocyanatononano onto glass substrate in vacuum. Deposition rates were improved at temperatures below 0 °C. After poling treatment films showed fairly large piezoelectric

activities. Additionally, a completely novel approach to piezoelectric polymers has been presented. This approach, consists in the synthesis of ordered piezoelectric polymer networks via crosslinking of liquid-crystalline monomers. The main goal in this approach is to achieve a polymer network which combines the long term stability of piezoelectric single-crystals with the ease of processability and fabrication of conventional polymers. [\[link\]](#) shows the schematic representation of this approach.



Scheme of a ordered piezoelectric networks via a liquid-crystalline monomer strategy. Adapted from D. L. Gin and B. C. Baxter, *Polymer Preprints*, 1996, **38**, 211.

Piezoelectric polymers are becoming increasingly important commercially because of their easier processability, lower cost, and higher impact resistance than ceramics, but the lack of high temperature stability and the absence of a solid understanding of the molecular level basis for the electrical properties are limitations. The requirements for strong piezoelectricity in a polymer are: the polymer chain has a larger resultant dipole moment normal to the chain axis; polymer crystallizes into a polar crystal with the polar axis perpendicular to the chain axis, has a high crystallinity and finally the polymer polar axis aligns easily in the thickness direction during poling.

Piezoelectric composites

Piezocomposites have been obtained by the combination of piezoelectric ceramics and polymers, the resulting material possesses both the high piezoelectric properties of ceramics and the processability of polymers. 1-3 type piezocomposites have found wide applications as medical and industrial ultrasonic transducers.

The current method for piezocomposite production is the dice-and-fill technique, which consists in cutting two sets of grooves in a block of piezoceramic at right angles each other, then a polymer is cast into these grooves and the solid ceramic base is ground off. Polishing and poling are the following steps in order to achieve the final thickness and properties. This method is expensive, time consuming and size limited.

As an alternative for the dice-and-fill technique, continuous green fibers obtained by the modified viscous-suspension-spinning process, can be bundled into a cottonball-like shape, then burned and sintered. The sintered bundle impregnated with epoxy resin can be sliced into discs and then polarized. Recent results have yielded 1-3 type composites with excellent piezoelectric properties.

On the other hand, an innovative process has been developed for $Sr_2(Nb_{0.5}Ta_{0.5})_2O_7/PVDF$ composites, in this new fabrication method, appropriate amounts of oxides are mixed, pressed and sintered. The porous resulting material is subsequently infiltrated with PVDF solution and then poled. This new method for composites preparation is simple and offers a lead-free alternative smart material.

Another kind of piezocomposites can be achieved by spinning films of piezoceramic onto metal alloys, such as TiNi. The resulting materials is a hybrid composite that can utilize the different active and adaptive

properties of the individual bulk materials. Due to the shape memory nature of TiNi, a possible application for this new heterostructures could be smart active damping of mechanical vibrations. DC sputtering and spin coating are the techniques necessary for the smart thin film TiNi/piezoelectric heterostructures fabrication. However, even though the films had a fine grain structure and high mechanical qualities, the ferroelectric properties were poor compared to literature values.

In the future, the properties of piezocomposites will be tailored, by varying the ceramic, the polymer and their relative proportions. Adjustments in the material properties will lead to fulfillment of the requirements for a particular device. [\[link\]](#) shows a comparison among piezoelectric ceramics, polymers and composites parameters where Z is the impedance, ϵ_{33}^t is the dielectrical constant, and ρ is the density.

Material parameter	Piezoceramics	Piezopolymers	Piezocomposites
k_t (%)	45 - 55	20 - 30	60 - 75
Z (10^6 Rayls)	20 - 30	1.5 - 4	4 - 20
$\epsilon_{33}^t/\epsilon_0$	200 - 5000	~10	50 - 2500
$\tan \gamma$ (%)	<1	1.5 - 5	<1
Q_m	10 - 1000	5 - 10	2 - 50
ρ (10^3 kg/m ³)	5.5 - 8	1 - 2	2 - 5

Parameter ranges for piezoelectric ceramics, polymers and composites.

Piezoelectric coatings.

Many potential applications exist which require film thickness of 1 to 30 μm . Some examples of these macroscopic devices include ultrasonic high frequency transducers, fiber optic modulators and for self controlled vibrational damping systems.

ZnO and PZT have been used for piezoelectric fiber optic phase modulators fabrication. The piezoelectric materials have been sputter deposited using dc magnetron source and multimagnetron sputtering systems. Coatings of 6 μm thick of ZnO and 0.5 μm of PZT are possible to achieve using these systems. However, thickness variation of approximately 15% occurs between the center and the end of ZnO coatings, results on affected modulation performance. Although PZT coatings achieved by sputtering posses uniformity and do not exhibit cracking, the PZT is only partially crystallized and it is actually a composite structure consisting of crystalline and amorphous material, diminishing the piezoelectric properties.

Sol-gel technique for thick PZT films have been developed. It is now possible to fabricate PZT sol-gel films of up to 60 μm . The electrical and piezoelectrical properties of the thick films reported are comparable with ceramic PZT.

Piezoelectric polymer coatings for high-frequency fiber-optic modulators have been also investigated. Commercial vinylidene fluoride and tetrafluoroethylene copolymer has been used. The advantage of using polymer coatings is that the polymer jacket (coating) can be easily obtained by melt extrusion on a single-

mode fiber. Thus, uniformity is easily achieved and surface roughness is not present. Furthermore, if annealing of the polymer is made prior poling, a high degree of crystallinity is enhanced, leading to better piezoelectric properties.

Bibliography

- R. N. Kleiman, *Mat. Res. Soc. Symp. Proc.*, 1996, **406**, 221.
- T. Yamamoto, *Jpn. J. Appl. Phys.*, 1996, **35**, 5104.
- Y. Yamashita, Y. Hosono, and N. Ichinose, *Jpn. J. Appl. Phys.*, 1997, **36**, 1141.
- I. Akimov and G. K. Savchuk, *Inorg. Mater.*, 1997, **33**, 638.
- L. Del Olmo and M. L. Calzada, *J. Non-Cryst. Solids*, 1990, **121**, 424.
- T. Nishi, K. Igarashi, T. Shimizu, K. Koumoto, and H. Yanagida, *J. Mater. Sci. Lett.*, 1989, **8**, 805.
- K. R. M. Rao, A. V. P. Rao, and S. Komarneni, *Mater. Lett.*, 1996, **28**, 463.
- K. Shimamura, H. Takeda, T. Kohno, and T. Fukuda, *J. Cryst. Growth*, 1996, **163**, 388.
- H. Takeda, K. Shimamura, T. Kohno, and T. Fukuda, *J. Cryst. Growth*, 1996, **169**, 503.
- Lee, T. Itoh and T. Suga, *Thin Solid Films*, 1997, **299**, 88.
- L. J. Mathias, D. A. Parrish, and S. Steadman, *Polymer*, 1994, **35**, 659.
- G. R. Fox, N. Setter, and H.G. Limberger, *J. Mater. Res.*, 1996, **11**, 2051.
- L. Gin and B. C. Baxter, *Polymer Preprints*, 1996, **38**, 211.

Formation of Silicon and Gallium Arsenide Wafers

Integrated circuits (ICs) and discrete solid state devices are manufactured on semiconductor wafers. The following focuses on the general principles and methods with regard to wafer formation.

Introduction

Integrated circuits (ICs) and discrete solid state devices are manufactured on semiconductor wafers. Silicon based devices are made on silicon wafers, while III-V (13-15) semiconductor devices are generally fabricated on GaAs wafers, however, for certain optoelectronic applications InP wafers are also used. The electrical and chemical properties of the wafer surface must be well controlled and therefore the preparation of starting wafers is a crucial portion of IC and device manufacturing. In order to obtain high fabrication yields and good device performance, it is very important that the starting wafers be of reproducibly high quality. For example, the front surface must be smooth and flat on both a macro- and microscale, because high-resolution patterns (lithography) are optically formed on the wafer. In principle, cutting a crystal into thin slices and polishing one side until all saw marks are removed and the surface appears smooth and glossy could produce a suitable wafer. However, due in part to the brittleness of Si and GaAs crystals, as well as the increasing requirements of wafer cleanliness and surface defect reduction with ever decreasing device geometries, a very complex series of processing steps are required to produce analytically clean, flat and damage-free wafer surfaces.

The following focuses on the general principles and methods with regard to wafer formation. Detailed formulas, recipes, and specific process parameters are not given as they vary considerably among different wafer producers. However, in general, techniques for fabrication of Si wafers have generally become standardized within the semiconductor industry. In contrast, GaAs wafer technology is less standardized, possibly due to either (a) the similarity to silicon practices or (b) the lower production volume of GaAs wafers. There are two general classes of processes in the methodology of making wafers: mechanical and chemical. As both Si and GaAs are brittle materials, the mechanical processes for their wafer fabrication are similar. However, the different chemistry of Si and GaAs require that the chemical processes be dealt with separately.

Wafer formation procedures

Each of the processing steps in the conversion of a semiconductor ingot (formed by Czochralski or Bridgeman growth) into a polished wafer ready for device fabrication, results in the removal of material from the original ingot; between $\frac{1}{3}$ and $\frac{1}{2}$ of the original ingot is sacrificed during processing. Methods for the removal of material from a crystal ingot are classified depending on the size of the particles being removed during the process. If the removed particles are much larger than atomic or molecular dimensions the process is described as being macro-scale. Conversely, if the material is removed atom-by-atom or molecule-by-molecule then the process is termed micro-scale. A further distinction between various types of processes is whether the removal occurs as a result of mechanical or

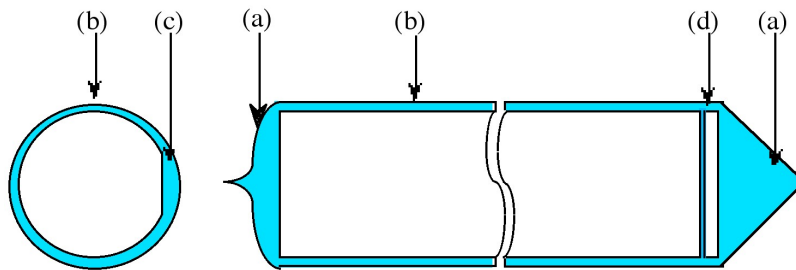
chemical processes. The formation of a finished wafer from a semiconductor ingot normally requires six machining (mechanical) operations, two chemical operations, and at least one polishing (chemical-mechanical) operation. Additionally, multiple inspection and evaluation steps are included in the overall process. A summary of the individual steps, and their functions, involved in wafer production is shown in [\[link\]](#).

Process	Type	Function
cropping	mechanical	removal of conical shaped ends and impure portions
grinding	mechanical	obtain precise diameter
orientation flatting	mechanical	identification of crystal orientation and dopant type
etching	chemical	removal of surface damage
wafering	mechanical	formation of individual wafers by cutting
heat treatment	thermal	annihilation of undesirable electronic donors
edge contouring	mechanical	provide radius on the edge of the wafer
lapping	mechanical	provides requisite flatness of the wafer
etching	chemical	removal of surface damage
polishing	mechano- chemical	provides a smooth (specular) surface
cleaning	chemical	removal of organics, heavy metals, and particulates

Summary of the process steps involved in semiconductor wafer production.

Crystal shaping

Although an as-grown crystal ingot is of high purity (99.9999%) and crystallinity, it does not have the sufficiently precise shape required for ready wafer formation. Thus, prior to slicing an ingot into individual wafers, several steps are needed. These operations required to prepare the crystal for slicing are referred to as crystal shaping, and are shown in [\[link\]](#).



Schematic representation of crystal shaping operations:

(a) remove crown and taper, (b) grind to required diameter, (c) grind flat, and (d) slice sample for measurements. Shaded area represents material removed.

Cropping

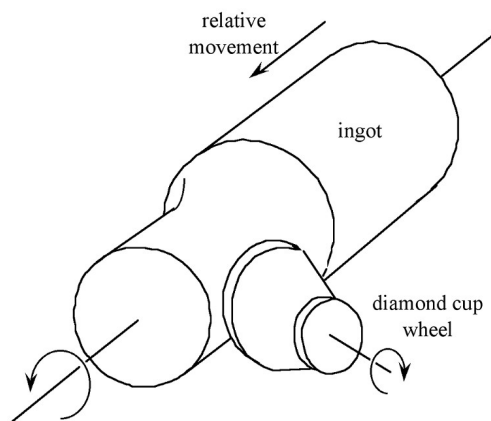
The as-grown ingots have conical shaped seed (top) and tang (bottom) ends that are removed using a circular diamond saw for ease of further manipulation of the ingot ([\[link\]](#)a). The cuttings are sufficiently pure that they are cleaned and recycled in the crystal growth operation. Portions of the ingot that fail to meet specifications of resistivity are also removed. In the case of silicon ingots these sections may be sold as metallurgical-grade silicon (MGS). Conversely, portions of the crystal that meet desired resistivity specifications may be preferentially selected. A sample slice is also cut to enable oxygen and carbon content to be determined; usually this is accomplished by Fourier transform infrared spectroscopic measurements (FT-IR). Finally, cropping is used to cut crystals to a suitable length to fit the saw capacity.

Grinding

The primary purpose of crystal grinding is to obtain wafers of precise diameter because the automatic diameter control systems on crystal growth equipment are not capable of meeting the tight wafer diameter specifications. In addition, crystals are seldom grown perfectly round in cross section. Thus, ingots are usually grown with a 1 - 2 mm allowance and reduced to the proper diameter by grinding [\[link\]](#)b.

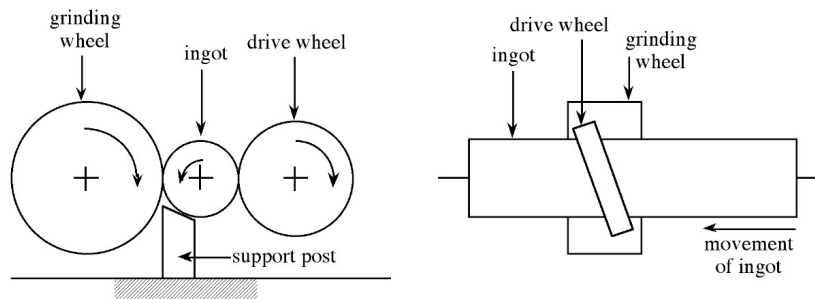
Crystal grinding is a straightforward process using an abrasive grinding wheel, however, it must be well controlled in order to avoid problems in subsequent operations. Excess chipping in wafering and lattice slip in thermal processing are problems often resulting from improper crystal grinding. Two methods are used for crystal grinding: (a) grinding on center and (b) centerless grinding.

[\[link\]](#) shows a schematic of the general set-up for grinding a crystal ingot on center. The crystal is supported at each end in a lathe-like machine. The rotating cutting tool, employing a water-based coolant, makes multiple passes down the rotating ingot until the requisite diameter is obtained. The center grinder can also be used for grinding the identification flats as well as providing a uniform ingot diameter. However, grinding the crystal on centers requires that the operator locate the crystal axis in order to obtain the best yield.



Schematic representation of grinding on center.

Centerless grinding eliminates the problems associated with locating the crystal center. The centerless method is superior for long crystals; however, a centerless grinder is much larger than a center grinder of the same diameter capacity. In centerless grinding the ingot is supported between two wheels, a grinding wheel and a drive wheel. A schematic of the centerless grinder is shown in [\[link\]](#). The axis of the drive wheel is canted with respect to that of the crystal ingot and the grinding wheel pushing the crystal ingot past the stationary (but rotating) grinding wheel, see [\[link\]](#)b.

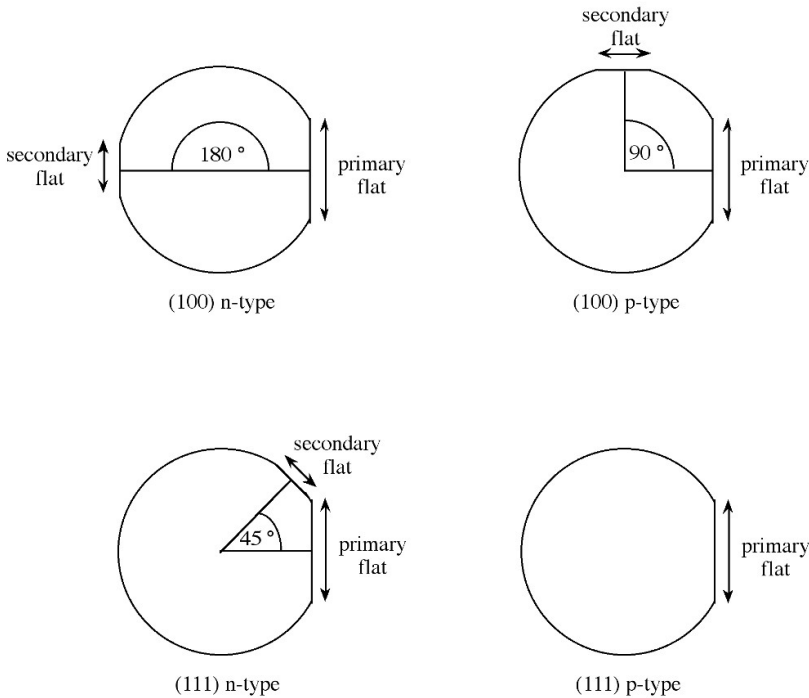


Schematic representation of centerless grinding viewed
(a) along and (b) perpendicular to the crystal axis.

Orientation/identification flats

Following grinding of the ingot to the desired diameter, one or two flats are ground along the length of the ingot. The identification flats (one or two) are ground lengthwise along the crystal according to the orientation and the dopant type. After grinding the crystal on centers the crystal is rotated to the proper orientation, then the wheel is positioned with its axis of rotation perpendicular to the crystal axis and moved along the crystal from end to end until the appropriate flat size is obtained. An optical or X-ray orientation fixture may be used in conjunction with the crystal mounting to facilitate the proper orientation of the crystal on the grinder.

The largest flat is called the primary flat ([link](#))c) and is parallel to one of the crystal planes, as determined by X-ray diffraction. The primary flat is used for automated positioning of the wafer during subsequent processing steps, e.g., lithographic patterning and dicing. Other smaller flats are called "secondary flats" and are used to identify the crystal orientation ($\langle 111 \rangle$ versus $\langle 100 \rangle$) and the material (n-type versus p-type). Secondary flats provide a quick and easy manner by which unknown wafers can be sorted. The flats shown schematically in [link](#) are located according to a Semiconductor Equipment and Materials Institute (SEMI[®]) standard and are ground to specific widths, depending upon crystals diameter. Notches are also used in place of the secondary flat; however, the relative orientations of the notch and primary flat with regard to crystal orientation and dopant are maintained.



SEMI locations for orientation/identification flats.

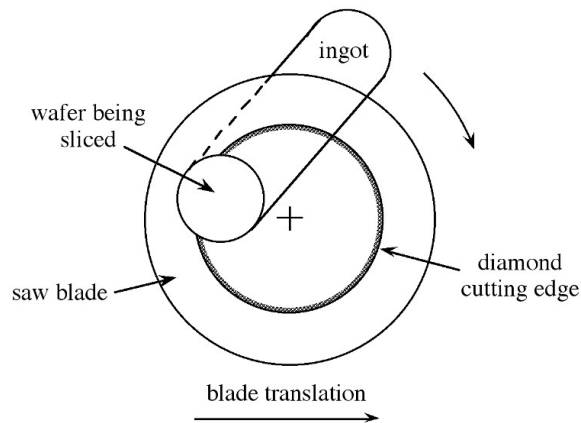
Etching

The cropping and grinding processes are performed with relatively coarse abrasive and consequently a great deal of subsurface damage results. Pits, chips, and cracks all contribute to stress in the cut wafer and provide nuclei for crack propagation at the edges of the finished wafer. If regions of stress are removed then cracks will no longer propagate, reducing exit chipping and wafer breakage during subsequent fabrication steps.

The general method for removing surface damage is to etch the crystal in a hot solution. The most common etchants for Si are based on the HNO_3 -HF system, in which etchant modifiers such as acetic acid also commonly used. In the case of GaAs HCl- HNO_3 is the appropriate system. These etchants selectively attack the crystal at the damaged regions. After etching, the crystal is transferred to the slicing preparation area.

Wafering

The purpose of wafering is to saw the crystal into thin slices with precise geometric dimensions. By far, the most common method of wafering semiconductor crystals is the use of an annular, or inner diameter (ID), diamond saw blade. A schematic diagram of ID slicing technology is shown in [\[link\]](#).








Schematic diagrams of ID slicing process.

The crystal, when it arrives at the sawing area, has been ground to diameter, flatted, and etched. In order to slice it, the crystal must be firmly mounted in such a way that it can be completely converted to wafers with minimum waste. The crystal is attached with wax or epoxy to a mounting block, which is usually cylindrical in shape and of the same diameter as the ingot. Also, a mounting beam (or strip) is attached along the length of the crystal at the breakout point of the saw blade. This reduces exit chipping (breakage that occurs as the blade exits the crystal at the end of a cut) and also provides support for the sawn wafer until it is retrieved. Graphite or phenolic resins are common materials for the mounting block and beams, although some success has been obtained in mounting ingots using hydraulic pressure. The saw blade is a thin sheet of stainless steel (325 μm), with diamond bonded to its inner edge. This blade is mounted on a drum that rotates at ca. 2000 rpm. Saw blades 58 cm (≈ 23 inches) in diameter with a 20 cm (8 inches) opening are common, however, as wafer sizes increase larger blades are employed: 30 cm (12 inches) wafers are now common for Si. The blade moves relative to the stationary crystal at a speed of 0.05 cm/s, and the cutting process is water-cooled. Thus, considering that wafers are sliced sequentially (one at a time), the overall process is very slow. A further problem is that the kerf loss (loss due to the width of the blade) results in approximately 1/3 of the material being lost as saw dust. Finally, the depth of the drum onto which the blade is attached limits the length of the ingot section that is accessible. In order to overcome this problem, another style of ID blade saw was developed in which the blade is mounted on an air bearing and is rotated by a belt drive. This allows the entire length of the crystal ingot to be sliced.

Both silicon and GaAs crystals are grown with either the crystallographic $\langle 100 \rangle$ or $\langle 111 \rangle$ direction parallel to the cylindrical axis of the crystal. Wafers may be cut either exactly perpendicular to the crystallographic axis or deliberately off-axis by several degrees. In order to obtain the proper wafer orientation, the crystal must be properly oriented on the saw. All production slicing machines have adjustments for orientation of the crystal; however, it is

usually necessary to check the orientation of the first slice in order to assure that all subsequent slices will be properly oriented.

Obvious variables introduced during the wafering process include: cutting rate, wheel speed, and coolant flow rate. However, the condition of the machines, such as alignment and vibration, is the most important variable followed by the condition of the blade. A deviated blade rim may cause taper, bow, or warp. [\[link\]](#) summarizes the types of deformations that can occur during wafering, their physical appearance and their characteristics.

Type of bow and warp	Surface appearance	Lattice curvature	Comments
	flat	flat	ideal
	curved	flat	
	curved	curved	
	flat	curved	
	curved	flat	slips

Deformed wafers and their characteristics.

Heat treatment

As-produced Czochralski grown crystals often have a level of oxygen impurity that may exceed the concentration of dopant in the semiconductor material (i.e., Si or GaAs). This oxygen impurity has a deleterious effect on the semiconductor properties, especially upon subsequent thermal processing, e.g., thermal oxide growth or epitaxial film growth by metal organic chemical vapor deposition (MOCVD). For example, when silicon crystals are heated

to about 450 °C the oxygen undergoes a transformation that causes it to behave as an electron donor, much like an n-type dopant. These oxygen donors, or "thermal donors", mask the true resistivity of the semiconductor because they either add additional carrier electrons to a n-type crystal or compensate for the positive holes in a p-type crystal. Fortunately, these thermal donors can be "annihilated" by heat treating the materials briefly in the range of 500 - 800 °C and then cooling quickly through the 450 °C region before donors can reform. In principle thermal donor annihilation can be performed on wafers at any time during their fabrication; however, it is usually best to perform the heat treatment immediately after wafering since sub-standard wafers may be rejected before additional processing steps are undertaken and thus limiting additional cost. Donor annihilation is a bulk effect, and therefore the thermal treatment can be performed in air, since any surface oxide that may form will be removed in subsequent lapping and polishing steps.

Lapping or grinding

The as-cut wafers vary sufficiently in thickness to require an additional operation, the slicing operation does not consistently produce the required flatness and parallelism required for many wafer specifications, see [\[link\]](#). Since conventional polishing does not correct variations in flatness or thickness, a mechanical two-sided lapping operation is performed. Lapping is capable of achieving very precise thickness uniformity, flatness and parallelism. Lapping also prepares the surface for polishing by removing the sub-surface sawing damage, replacing it with a more uniform and smaller lapping damage.

The process used for lapping semiconductor wafers evolved from the optical lens manufacturing industry using principles developed over several hundred years. However, as the lens has a curved surface and the wafers are flat, the equipment for lapping wafers is mechanically simpler than lens processing machines. The simplest double-side lapping machine consists of two very flat counter-rotating plates, carriers to hold and move the wafers between the plates, and a device to feed abrasive slurry steadily between the plates. The abrasive is typically a 9 µm Al₂O₃ grit. Commercial abrasives are suspended in water or glycerin with proprietary additives to assist in suspension and dispersion of the particles, to improve the flow properties of the slurry, and to prevent corrosion of the lapping machine. Hydraulics or an air cylinder applies lapping pressure with low starting pressure for 2 to 5 minutes, which is then increased through most of the process. The completion of lapping may be determined by elapsed time or by an external thickness sensing device. The finished process gives a wafer with a surface uniform to within 2 µm. Approximately 20 µm per side is removed during the lapping process.

Although lapping would appear to be simple in concept, the successful implementation of a production lapping operation requires the development of a technique and experience to achieve acceptable quality with good yields. Small adjustments to the rotation rates of the plates and carriers will cause the plates to wear concave, convex or flat.

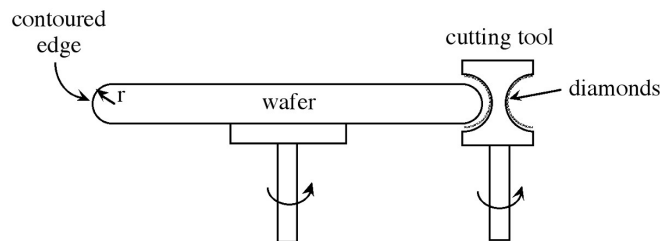
As lapping is a messy process, various efforts have been made to avoid it or to substitute an alternative process. The most likely approach at present is grinding, in which the wafer is held on a vacuum chuck and a series of progressively finer diamond wheels is moved over

the wafer while it is rotated on a turn table. Grinding gives a clearer surface than lapping, however, only one side may be ground at a time and the resulting flatness is not as good as that obtained by lapping.

Edge contouring

The rounding of the edge of the wafer to a specific contour is a fairly recent development in the technology of wafer preparation. It was known by the early seventies that a significant number of device yield problems could be traced to the physical condition of the wafer edge. An acute edge affects the strength of the wafer due to: stress concentration, and a lowering of its resistance to thermal stress, as well as being the source of particle chip, breakage, and lattice damage. In addition, the particles originating from the chipped edges can, if present on the wafer surface, add to the defect density (D_0) of the IC process reducing fabrication yield. Further problems associated with a square edge include the build-up of photoresist at the wafer edge. The solution to these process problems is to provide a contoured edge with a defined radius (r).

Chemical etching of wafers results in a degree of edge rounding, but it is difficult to control. Thus, mechanical edge contouring has been developed and the result has been a dramatic improvement in yields in downstream wafer processing. Losses due to wafer breakage are also reduced. The edge contouring process is usually performed in cassette-fed high speed equipment, in which each wafer is rotated rapidly against a shaped cutting tool ([link](#)).



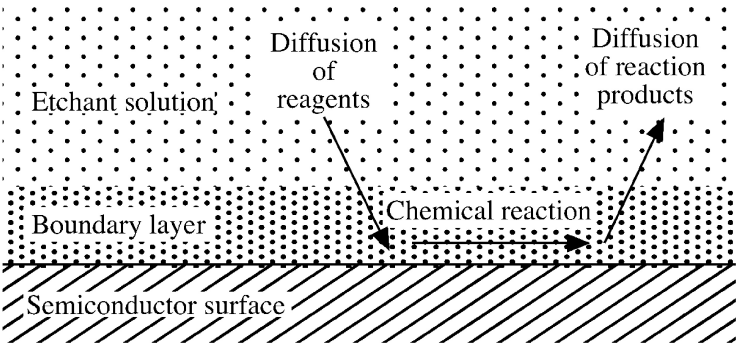
Schematic illustration of edge contouring.

Etching

The mechanical processes described above to shape the wafer leave the surface and edges damaged and contaminated. The depth of the work damage depends on the specific process, however, 10 μm is typical. Such damage is readily removed by chemical etching. Etching is used at multiple points during the fabrication of a semiconductor device. The discussion below is limited to etches suitable for wafer fabrication, i.e., non-selective etching of the entire wafer surface.

Wet chemical etching

The wet chemical etching of any material can be considered to involve three steps: (a) transportation of the reactants to the surface, (b) reaction at the surface, and (c) movement of the reaction products into the etchant solution ([\[link\]](#)). Each of these may be the rate limiting step and thus control the etch rate and uniformity. This effect is summarized in [\[link\]](#).



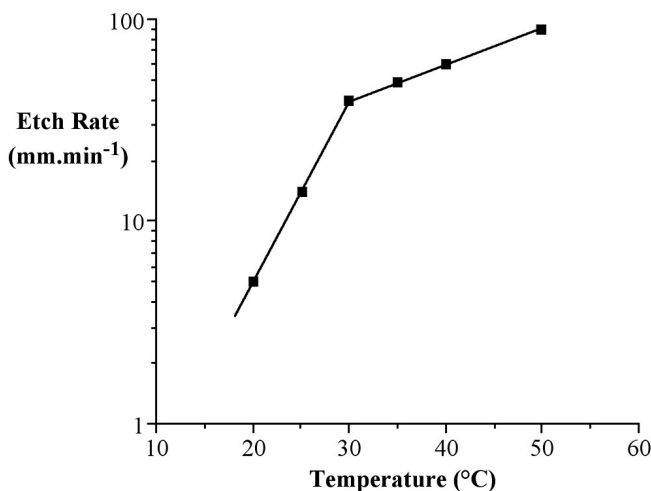
Schematic representation of the three steps involved in wet chemical etching: (i) diffusion of the chemical etch reagents through the boundary layer, (ii) chemical reaction at the surface, and (iii) diffusion of the reaction products into the etch solution through the boundary layer.

Rate limiting step	Etching rate	Results	Comments
Diffusion of reagent to the surface	slow	etching(anisotropic)	enhanced surface roughness
Reaction at semiconductor surface	fast	polishing(isotropic)	ideal
Diffusion of reaction products from the surface	slow	polishing(isotropic)	reaction product

			remains on surface
--	--	--	--------------------

Effects of rate limiting step in semiconductor etching.

An etchant that is limited by the rate of reaction at the surface will tend to enhance any surface features and promote surface roughness due to preferential etching at defects (anisotropic). In contrast, if the etch rate is limited by the diffusion of the etchant reagent through a stagnant (dead) boundary layer near the surface, then the etch will result in uniform polishing and the surface will become smooth (isotropic). If removal of the reaction products is rate limiting then the etch rate will be slow because the etch equilibrium will be shifted towards the reactants. In the case of an individual etchant reaction, the rate determining step may be changed by rapid stirring to aid removal of reaction products, or by increasing the temperature of the etch solution, see [\[link\]](#). The exact etching conditions are chosen depending on the application. For example, dilute high temperature etches are often employed where the etch damage must be minimized, while cooled etches can be used where precise etch control is required.



Typical etch rate versus temperature plot for a mixture of HF (20%), nitric acid (45%), and acetic acid (35%).

Traditionally mixtures of hydrofluoric acid (HF), nitric acid (HNO₃) and acetic acid (MeCO₂H) have been used for silicon, but alkaline etches using potassium hydroxide (KOH) or sodium hydroxide (NaOH) solutions are increasingly common. Similarly, gallium arsenide etches may be either acidic or basic, however, in both cases the etches are oxidative due to the use of hydrogen peroxide. A wide range of chemical reagents are commercially available in "transistor grade" purity and these are employed to minimize contamination of the

semiconductor. Deionized water is commonly used as a diluent for each of these reagents and the concentration of commonly used aqueous reagents is given in [\[link\]](#).

Reagent	Weight %	Reagent	Weight %
HCl	37	HF	49
H ₂ SO ₄	98	H ₃ PO ₄	85
HNO ₃	79	HClO ₄	70
MeCO ₂ H	99	H ₂ O ₂	30
NH ₄ OH	29		

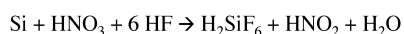
Weight percent concentration of commonly used concentrated aqueous reagents.

The equipment used for a typical etchant process includes an acid (or alkaline) resistant tank, which contains the etchant solution and one or more positions for rinsing the wafers with deionized water. The process is batch in nature involving tens of wafers and the best equipment provides a means of rotating the wafers during the etch step to maintain uniformity. In order to assure the removal of all surface damage, substantial over-etching is performed. Thus, the removal of 20 µm from each side of the wafer is typical. Etch times are usually several minutes per batch.

Etching silicon

The most commonly used etchants for silicon are mixtures of hydrofluoric acid (HF) and nitric acid (HNO₃) in water or acetic acid (MeCO₂H). The etching involves a reduction-oxidation (redox) reaction, followed by dissolution of the reaction products. In the HF-HNO₃ system the HNO₃ oxidizes the silicon and the HF removes the reaction products from the surface. The overall reaction is:

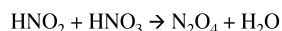
Equation:



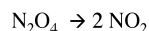
The oxidation reaction involves the oxidation of Si⁰ to Si⁴⁺, and it is auto-catalytic in that the reaction product promotes the reaction itself. The initial step involves trace impurities of

HNO₂ in the HNO₃ solution, [\[link\]](#), which react to liberate nitrogen dioxide (NO₂), [\[link\]](#).

Equation:

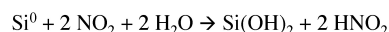


Equation:

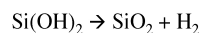


The nitrogen dioxide oxidizes the silicon surface in the presence of water, resulting in the formation of Si(OH)₂ and the reformation of HNO₂, [\[link\]](#). The Si(OH)₂ decomposes to give SiO₂, [\[link\]](#). Since the reaction between HNO₂ and HNO₃, [\[link\]](#), is rate limiting, an induction period is observed. However, this is overcome by the addition of NO₂⁻ ions in the form of [NH₄][NO₂].

Equation:

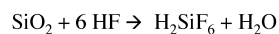


Equation:



The final step of the etch process is the dissolution of the SiO₂ by HF, [\[link\]](#). Stirring serves to remove the soluble products from the reaction surface. The role of the HF is to act as a complexing reagent, and thus the reaction shown in [\[link\]](#) is known as a complexing reaction. The formation of water as a reaction product requires that acetic acid be used as a diluent (solvent) to ensure better control.

Equation:



The etching reaction is highly dependent on the relative ratios of the etchant reagents. Thus, if an HF-rich solution is used, the reaction is limited by the oxidation step, [\[link\]](#), and the etching is anisotropic, since the oxidation reaction is sensitive to doping, crystal orientation, and defects. In contrast, the use of a HNO₃-rich solution produces isotropic etching since the dissolution process is rate limiting ([\[link\]](#)). The reaction of HNO₃-rich solutions has been found to be diffusion-controlled over the temperature range 20 - 50 °C ([\[link\]](#)), and is therefore commonly employed for removing work damage produced during wafer fabrication. The boundary layer thickness ([\[link\]](#)) and therefore the dimensional control over the wafer is controlled by the rotation rate of the wafers. A common etch formulation is a 4:1:3 mixture of HNO₃ (79%), HF (49%), and MeCO₂H (99%). There are some etchant

formulations that are based on alternative (or additional) oxidizing agents, such as: Br₂, I₂, and KMnO₄.

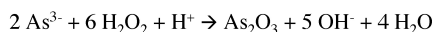
Alkaline etching (KOH/H₂O or NaOH/H₂O) is by nature anisotropic and the etch rate depends on the number of dangling bonds which in turn are dependent on the surface orientation. Since etching is reaction rate limited no rotation of the wafers is necessary and excellent uniformity over large wafers is obtained. Alkaline etchants are used with large wafers where dimensional uniformity is not maintained during lapping. A typical formulation uses KOH in a 45% weight solution in H₂O at 90 °C.

Etching gallium arsenide

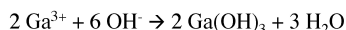
Although a wide range of etches have been investigated for GaAs, few are truly isotropic. This is because the surface activity of the (111) Ga and (111) As faces are very different. The As rich face is considerably more reactive than the Ga rich face, thus under identical conditions it will etch faster. As a result most etches give a polished surface on the As face, but the Ga face tends to appear cloudy or frosted due to the highlighting of surface features and crystallographic defects.

As with silicon the etch systems involve oxidation and complexation. However, in the case of GaAs the gallium is already fully oxidized (formally Ga³⁺), thus, it is the arsenic (formally the arsenide ion, As³⁻ that is oxidized by a suitable oxidizing agent (e.g., H₂O₂) to the soluble oxide, As₂O₃, [\[link\]](#). The gallium ions form the oxide Ga₂O₃ via the hydroxide, [\[link\]](#). Both oxides are soluble in acid solutions, resulting in their removal from the surface.

Equation:



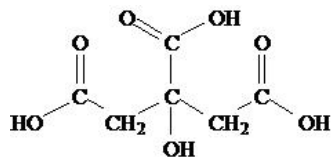
Equation:



The peroxide based oxidative etches for GaAs are divided into acidic and basic etches. The composition and application of some of these systems are summarized in [\[link\]](#). The most widely used of these is H₂SO₄/H₂O₂/H₂O and is referred to as Caro's acid. The high viscosity of H₂SO₄ results in diffusion-limited etching with high acid concentrations. Etches with low acid concentrations tend to be anisotropic. Phosphoric acid (H₃PO₄) or citric acid ([\[link\]](#)) may be exchanged for sulfuric acid (H₂SO₄). Replacement of the acid component with bases such as NH₄OH or NaOH can result in near to truly isotropic etchants, although certain combinations can result in strong anisotropy.

Formulation	Volume ratio	(100) etch rate ($\mu\text{m}/\text{min}$)	(110) etch rate ($\mu\text{m}/\text{min}$)	(111)As etch rate ($\mu\text{m}/\text{min}$)	(111)Ga etch rate ($\mu\text{m}/\text{min}$)
$\text{H}_2\text{SO}_4/\text{H}_2\text{O}_2/\text{H}_2\text{O}$	8:1:1	1.5	1.5	1.5	0.8
$\text{H}_2\text{SO}_4/\text{H}_2\text{O}_2/\text{H}_2\text{O}$	1:8:1	8.0	8.0	12.0	3.0
$\text{H}_2\text{SO}_4/\text{H}_2\text{O}_2/\text{H}_2\text{O}$	3:1:50	0.8	0.8	0.8	0.4
citric acid/ $\text{H}_2\text{O}_2/\text{H}_2\text{O}$	1:1:1	0.6	0.6	0.6	0.4
$\text{NH}_4\text{OH}/\text{H}_2\text{O}_2/\text{H}_2\text{O}$	1:700	0.3	0.3	0.3	0.3
$\text{NaOH}/\text{H}_2\text{O}_2/\text{H}_2\text{O}$	1:0.76	0.2	0.2	0.2	0.2

The composition and application of selected etch systems for GaAs.



Structure of citric acid.

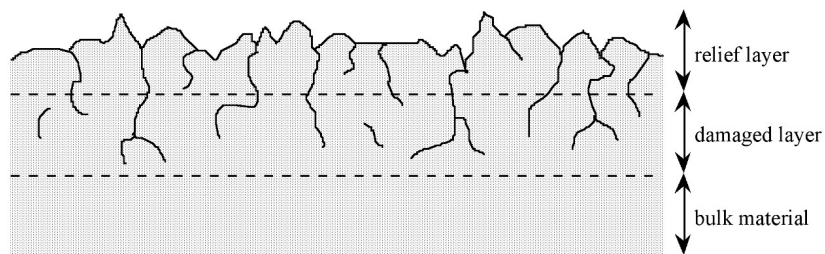
One of the earliest etching systems for GaAs is based on the use of a dilute (ca. 0.05 vol.%) solution of bromine (Br_2) in ethanol. The Br_2 acts as the oxidant, resulting in the formation of soluble bromides. The etch rate of this system is different for different crystallographic planes, i.e., the etch rates for the (111) As, (100), and (111) Ga faces are in the ratio 6:5:1, although more uniform etch rates are observed with high Br_2 concentrations (ca. 10 vol.%). These higher concentration solutions are used for the removal of damage due to cutting with the saw.

Polishing

The purpose of polishing is to produce a smooth, specular surface on which device features can be defined by lithography. In order to allow for very large scale integration (VLSI) or ultra large scale integration (ULSI) fabrication the wafer must have a surface with a high degree of flatness. Variations less than 5 to 10 μm across the wafer diameter are typical flatness specifications. In addition, given the preceding steps, wafer polishing must not leave residual contamination or surface damage. The techniques of wafer polishing are derived from the glass lens industry, with some important modifications that have been developed to meet the special requirements of the microelectronics industry.

Differences between polishing and lapping

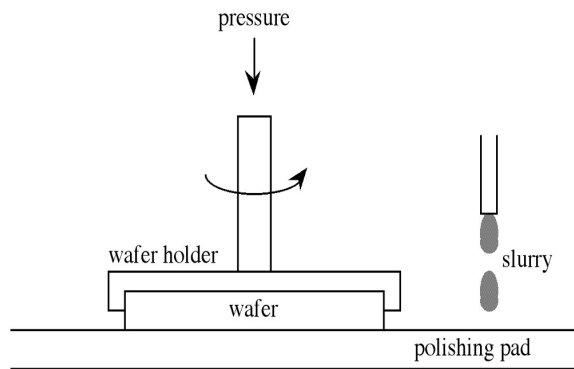
If the surface of a wafer that has undergone lapping (or grinding) is examined with an electron microscope, cracks, ridges and valleys are observed. The top "relief layer" consists of peaks and valleys. Below this layer is a damaged layer characterized by microcracks, dislocations, slip and stress. [\[link\]](#) shows a schematic representation of the abraded surface. Both of these layers must be removed completely prior to further fabrication. Decreasing the particle size of the abrasive during lapping only decreases the scale of the damage, but does not eliminate it entirely. In fact this surface damage is a characteristic of the brittle fracture of single crystal Si and GaAs, and occurs because during lapping the abrasive grains are moved across the surface under a pressure beyond that of the fracture strength of the wafer materials (Si or GaAs). In contrast to the mechanical abrasion employed in lapping, polishing is a mechano-chemical process during which brittle fracture does not occur. A polished wafer does not display any evidence of a relief surface such as that produced by lapping, even at highest resolution electron microscope.



Schematic representation of a cross sectional view of an abraded wafer surface prior to polishing.

Process of Polishing

[\[link\]](#) shows a schematic of the polishing process. Polishing may be conducted on single wafers or as a batch process depending on the equipment employed. Single wafer polishing is preferred for larger wafers and allows for better surface flatness. In both processes, wafers are mounted onto a fixture, by either wax or a composite Felx-Mount™, and pressed against the polishing pad. The polishing pad is usually made from an artificial fabric such as polyester felt-polyurethane laminate. Polishing is accomplished by a mechano-chemical process in which aqueous polishing slurry is dripped onto the polishing pad, see [\[link\]](#). The polishing slurry performs both a chemical and mechanical process, and consists of fine silica (SiO_2) particles (100 Å diameter) and an oxidizing agent. Aqueous sodium hydroxide (NaOH) is used for Si, while aqueous sodium chlorate (NaOCl) is preferred for GaAs. Suspending agents are usually added to prevent settling of the silica particles. Under the heat caused by the friction of the wafer on the polishing pad the wafer surface is oxidized, which is the chemical step, while in the mechanical step the silica particles in the slurry abrade the oxidized surface away.



Schematic representation of the wafer polishing process.

In order to achieve a reasonable rate of removal of the relief and damaged layers and still obtain the highest quality surface, the polishing is done in two steps, stock removal and haze removal. The former is carried out with a higher concentration slurry and may proceed for about 30 minutes at a removal rate of 1 $\mu\text{m}/\text{min}$. Haze removal is performed with a very dilute slurry, a softer pad with a reaction time of about 5 to 10 minutes, during which the total amount of material removed is only about 1 μm . Due to the active chemical reaction between the wafer and the polishing agent, the wafers must be rinsed in deionized water immediately after polishing to prevent haze or stains from reforming.

There are many variables that will influence the rate and quality of polishing. High pressure results in a higher polishing rate, but excessive pressure may cause non-uniform polishing, excessive heat generation and fast pad wear. The rate of polishing is increased with higher temperatures but this may also lead to haze formation. High wheel speeds accelerate the

polishing rate but can raise the temperature and also results in problems in maintaining a uniform flow of slurry across the pad. Dense slurry concentrations increase the polishing rate but are more costly. The pH of the slurry solution can also affect the polishing rate, for example the polishing rate of Si gradually increases with increased pH (higher basicity) until a pH of about 12 where a dramatic decrease is observed. In general, the optimum polishing process for a given facility depends largely upon the interplay of product specification, yields, cost, and quality considerations and must be developed uniquely. The wafer polishing process does not improve the wafer flatness and, at best, polishing will not degrade the wafer flatness achieved in the lapping operation.

Cleaning

During the processes described above, semiconductor wafers are subjected to physical handling that leads to significant contamination. Possible sources of physical contamination include:

- a. airborne bacteria,
- b. grease and wax from cutting oils and physical handling,
- c. abrasive particulates (usually, silica, silicon carbide, alumina, or diamond dust) from lapping, grinding or sawing operations,
- d. plasticizers which are derived from containers and wrapping in which the wafers are handled and shipped.

Chemical contamination may also occur as a result of improper cleaning after etch steps. Light-metal (especially sodium and potassium) species may be traced to impurities in etchant solutions and are chemisorbed on to the surface where they are particularly problematical for metal oxide semiconductor (MOS) based devices, although higher levels of such impurities are tolerable for bipolar devices. Heavy metal impurities (e.g., Cu, Au, Fe, and Ag) are usually caused by electrodeposition from etchant solutions during fabrication. While wafers are cleaned prior to shipping, contamination accumulated during shipping and storage necessitates that all wafers be subjected to scrupulous cleaning prior to fabrication. Furthermore, cleaning is required at each step during the fabrication process. Although wafer cleaning is a vital part of each fabrication step, it is convenient to discuss cleaning within the general topic of wafer fabrication.

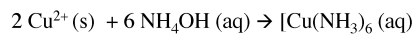
Cleaning silicon

The first step in cleaning a Si wafer is removal of all physical contaminants. These contaminants are removed by rinsing the wafer in hot organic solvents such as 1,1,1-trichloroethane (Cl_3CH_3) or xylene ($\text{C}_6\text{H}_4\text{Me}_2$), accompanied by mechanical scrubbing, ultrasonic agitation, or compressed gas jets. Removal of the majority of light metal contaminants is accomplished by rinsing in hot deionized water, however, complete removal requires a further more aggressive cleaning process. The most widely used cleaning method

in the Si semiconductor industry is based on a two step, two solution sequence known as the “RCA Cleaning Method”.

The first solution consists of H₂O-H₂O₂-NH₄OH in a volume ratio of 5:1:1 to 7:2:1, which is used to remove organic contaminants and heavy metals. The oxidation of the remaining organic contaminants by the hydrogen peroxide (H₂O₂) produces water soluble products. Similarly, metal contaminants such as cadmium, cobalt, copper, mercury, nickel, and silver are solubilized by the NH₄OH through the formation of soluble amino complexes, e.g., [\[link\]](#).

Equation:



The second solution consists of H₂O-H₂O₂-HCl in a 6:1:1 to 8:2:1 volume ratio and removes the Group I(1), II(2) and III(13) metals. In addition, the second solution prevents re-deposition of the metal contaminants. Each of the washing steps is carried out for 10 - 20 min. at 75 - 85 °C with rapid agitation. Finally, the wafers are blown dry under a stream of nitrogen gas.

Cleaning GaAs

In principle GaAs wafers may be cleaned in a similar manner to silicon wafers. The first step involves successive cleaning with hot organic solvents such as 1,1,1-trichloroethane, acetone, and methanol, each for 5-10 minutes. GaAs wafers cleaned in this manner may be stored under methanol for short periods of time.

Most cleaning solutions for GaAs are actually etches. A typical solution is similar to the second RCA solution and consists of an 80:10:1 ratio of H₂O-H₂O₂-HCl. This solution is generally used at elevated temperatures (70 °C) with short dip times since it has a very fast etch rate (4.0 μm/min).

Measurements, inspections and packaging

Quality control measurements of the semiconductor crystal and subsequent wafer are performed throughout the process as an essential part of the fabrication of wafers. From crystal and wafer shaping through the final wafer finishing steps, quality control measurements are used to ensure that the materials meets customer specifications, and that problems can be corrected before they create scrap material and thus avoid further processing of reject material. Quality control measurements can be broadly classified into mechanical, electrical, structural, and chemical.

Mechanical measurements are concerned with the physical dimensions of the wafer, including: thickness, flatness, bow, taper and edge contour. Electrical measurements usually include: resistivity and lateral resistivity gradient, carrier type and lifetime. Measurements

giving information on the perfection of the semiconductor crystal lattice are classified in the structural category and include: testing for stacking faults, and dislocations. Routine chemical measurements are limited to the measurement of dissolved oxygen and carbon by Fourier transform infrared spectroscopy (FT-IR). Finished wafers are individually marked for the purpose of identification and traceability. Packaging helps protect the finished wafers from contamination during shipping and storage.

Industry standards defining in detail how quality control measurements are to be made and determining the acceptable ranges for measured values have been developed by the American Society of Testing Materials (ASTM) and the Semiconductor Equipment and Materials Institute (SEMI).

Bibliography

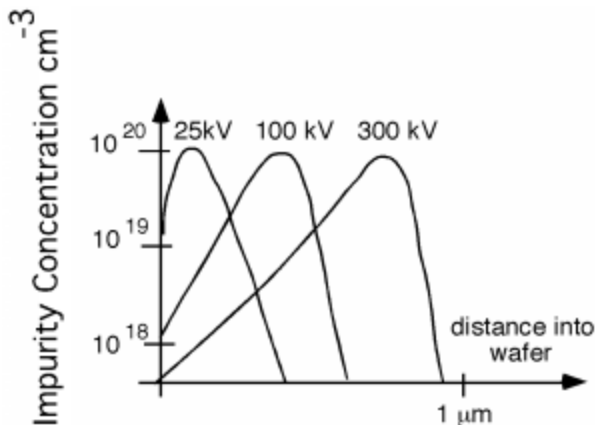
- A. C. Bonora, *Silicon Wafer Process Technology: Slicing, Etching, Polishing*, Semiconductor Silicon 1977, Electrochem. Soc., Pennington, NJ (1977).
- L. D. Dyer, in *Proceeding of the low-cost solar array wafering workshop 1981*, DoE-JPL-21012-66, Jet Propulsion Lab., Pasadena CA (1982).
- J. C. Dymant and G. A. Rozgonyi, *J. Electrochem. Soc.*, 1971, **118**, 1346.
- H. Gerischer and W. Mindt, *Electrochem. Acta*, 1968, **13**, 1329.
- P. D. Green, *Solid State Electron.*, 1976, **19**, 815.
- C. A. Harper and R. M. Sompson, *Electronic Materials & Processing Handbook*, McGraw Hill, New York, 2nd Edition.
- S. Iida and K. Ito, *J. Electrochem. Soc.*, 1971, **118**, 768.
- W. Kern, *J. Electrochem. Soc.*, 1990, **137**, 1887.
- Y. Mori and N. Watanabe, *J. Electrochem. Soc.*, 1978, **125**, 1510.
- D. L. Partin, A. G. Milnes, and L. F. Vassamillet, *J. Electrochem. Soc.*, 1979, **126**, 1581.
- D. W. Shaw, *J. Electrochem. Soc.*, 1966, **113**, 958.
- F. Snimura, *Semiconductor Silicon Crystal Technology*, Academic Press, New York (1989).
- D. R. Turner, *J. Electrochem. Soc.*, 1960, **107**, 810.

Doping

Starting with a prepared, polished wafer then how do we get an integrated circuit? We will focus on the CMOS process, described in the last chapter. Let's assume we have wafer which was doped during growth so that it has a background concentration of acceptors in it (i.e. it is p-type). Referring back to [CMOS Logic](#), you can see that the first thing we need to build is a n-tank or moat. In order to do this, we need some way in which to introduce additional impurities into the semiconductor. There are several ways to do this, but current technology relies almost exclusively on a technique called **ion implantation**. A diagram of an ion-implanter is shown in the [figure in the previous section](#). An ion implanter uses a dopant source gas, ionizes it, and drives the ions into the wafer. The dopant gas is ionized and the resultant charged ions are accelerated through a magnetic field, where they are mass-analyzed. The vertical magnetic field causes the beam of ions to spread out, according to their mass. A thin aperture selects the ions of interest, and lets them pass, blocking all the others. This makes sure we are only implanting the ion we want, and in fact, even selects for the proper isotope! The ionized atoms are then accelerated through several tens to hundreds of kV, and then deflected by an electric field, much like in an oscilloscope CRT. In fact, most of the time the ion beam is "rastered" across the surface of the silicon wafer. The ions strike the silicon wafer and pass into its interior. A measurement of the current flow in the system and its integral, is a measure of how much dopant was deposited into the wafer. This is usually given in terms of the number of dopant $\frac{\text{atoms}}{\text{cm}^2}$ to which the wafer has been exposed.

After the atoms enter the silicon, they interact with the lattice, creating defects, and slowing down until finally they stop. Typical atomic distributions, as a function of implant voltage are show in [\[link\]](#) for implantation into amorphous silicon. When implantation is done on single crystal material, channeling, the improved mobility of an ion down the "hallway" of a given lattice direction, can skew the impurity distribution significantly. Just slight changes of less than a degree can make big differences in how the impurity atoms are finally distributed in the wafer. Usually, the operator of the implant machine purposely tilts the wafer a few

degrees off normal to the beam in order to arrive at more reproducible results.



Implant distribution with
acceleration energy

As you might expect, shooting 100 kV ions at a silicon wafer probably does quite a bit of damage to the crystal structure. Not only that, but just having, say boron, in your wafer does not mean you are going to have holes. For the boron to become "electrically active" - that is to act as an acceptor - it has to reside on a silicon lattice site. Even if the boron atom does, somehow, end up on an actual lattice site when it stops crashing around in the wafer, the many defects which have been created will act as deep traps. Thus, the hole which is formed will probably be caught at a trap site and will not be able to contribute to electrical conductivity in the wafer anyway. How can we fix this situation? If we carefully heat up the wafer, we can cause the atoms in the crystal to shake around, and if we do it right, they all get back where they belong. Not only that, but the newly added impurities end up on lattice sites as well! This step is called **annealing** and it does just what it is supposed to. Typical temperatures and times for such an anneal are 500 to 1000°C for 10 to 30 minutes.

Something else occurs during the anneal step however. We have just added, by our implantation step, impurities with a fairly tight distribution as shown in [\[link\]](#). There is an obvious gradient in impurity distribution, and if there

is a gradient, than things may start moving around by diffusion, especially at elevated temperatures.

Applications for Silica Thin Films

Introduction

While the physical properties of silica make it suitable for use in protective and optical coating applications, the biggest application of insulating SiO₂ thin films is undoubtedly in semiconductor devices, in which the insulator performs a number of specific tasks, including: surface passivation, field effect transistor (FET) gate layer, isolation layers, planarization and packaging.

The term insulator generally refers to a material that exhibits low thermal or electrical conductivity; electrically insulating materials are also called dielectrics. It is in regard to the high resistance to the flow of an electric current that SiO₂ thin films are of the greatest commercial importance. The dielectric constant (ϵ) is a measure of a dielectric materials ability to store charge, and is characterized by the electrostatic energy stored per unit volume across a unit potential gradient. The magnitude of ϵ is an indication of the degree of polarization or charge displacement within a material. The dielectric constant for air is 1, and for ionic solids is generally in the range of 5 - 10. Dielectric constants are defined as the ratio of the material's capacitance to that of air, i.e., [\[link\]](#). The dielectric constant for silicon dioxide ranges from 3.9 to 4.9, for thermally and plasma CVD grown films, respectively.

Equation:

$$\epsilon = C_{\text{material}}/C_{\text{air}}$$

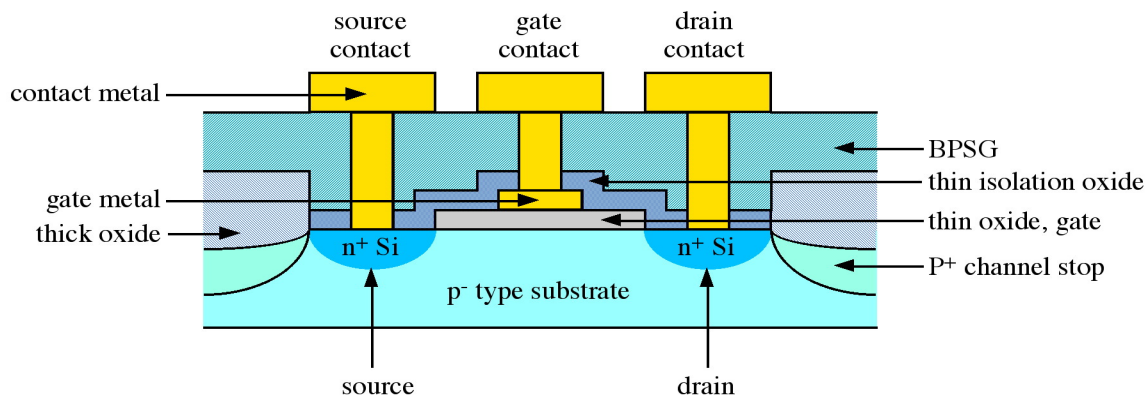
An insulating layer is a film or deposited layer of dielectric material separating or covering conductive layers. Ideally, in these application an insulating material should have a surface resistivity of greater than 10¹³ Ω/cm² or a volume resistivity of greater than 10¹¹ Ω.cm. However, for some applications, lower values are acceptable; an electrical insulator is generally accepted to have a resistivity greater than 10⁵ Ω.cm. CVD SiO₂ thin films have a resistivity of 10⁶ - 10¹⁶ Ω.cm, depending on the film growth method.

As a consequence of its dielectric properties SiO_2 , and related silicas, are used for isolating conducting layers, to facilitate the diffusion of dopants from doped oxides, as diffusion and ion implantation masks, capping doped films to prevent loss of dopant, for gettering impurities, for protection against moisture and oxidation, and for electronic passivation. Of the many methods used for the deposition of thin films, chemical vapor deposition (CVD) is most often used for semiconductor processing. In order to appreciate the unique problems associated with the CVD of insulating SiO_2 thin films it is worth first reviewing some of their applications. Summarized below are three areas of greatest importance to the fabrication of contemporary semiconductor devices: isolation and gate insulation, passivation, and planarization.

Device isolation and gate insulation

A microcircuit may be described as a collection of devices each consisting of "an assembly of active and passive components, interconnected within a monolithic block of semiconducting material". Each device is required to be isolated from adjacent devices in order to allow for maximum efficiency of the overall circuit. Furthermore within a device, contacts must also be electrically isolated. While there are a number of methods for isolating individual devices within a circuit (reverse-biased junctions, mesa isolation, use of semi-insulating substrates, and oxide isolation), the isolation of the active components in a single device is almost exclusively accomplished by the deposition of an insulator.

In [\[link\]](#) is shown a schematic representation of a silicon MOSFET (metal-oxide-semiconductor field effect transistor). The MOSFET is the basic component of silicon-CMOS (complimentary metal-oxide-semiconductor) circuits which, in turn, form the basis for logic circuits, such as those used in the CPU (central processing unit) of a modern personal computer. It can be seen that the MOSFET is isolated from adjacent devices by a reverse-biased junction (p^+ -channel stop) and a thick oxide layer. The gate, source and drain contact are electrically isolated from each other by a thin insulating oxide. A similar scheme is used for the isolation of the collector from both the base and the emitter in bipolar transistor devices.



Schematic diagrams of a Si-MOSFET (metal-oxide-semiconductor field effect transistor).

As a transistor, a MOSFET has many advantages over alternate designs. The key advantage is low power dissipation resulting from the high impedance of the device. This is a result of the thin insulation layer between the channel (region between source and drain) and the gate contact, see [\[link\]](#). The presence of an insulating gate is characteristic of a general class of devices called MISFETs (metal-insulator-semiconductor field effect transistor). MOSFETs are a subset of MISFETs where the insulator is specifically an oxide, e.g., in the case of a silicon MISFET device the insulator is SiO_2 , hence MOSFET. It is the fabrication of MOSFET circuits that has allowed silicon technology to dominate digital electronics (logic circuits). However, increases in computing power and speed require a constant reduction in device size and increased complexity in device architecture.

Passivation

Passivation is often defined as a process whereby a film is grown on the surface of a semiconductor to either (a) chemically protect it from the environment, or (b) provide electronic stabilization of the surface.

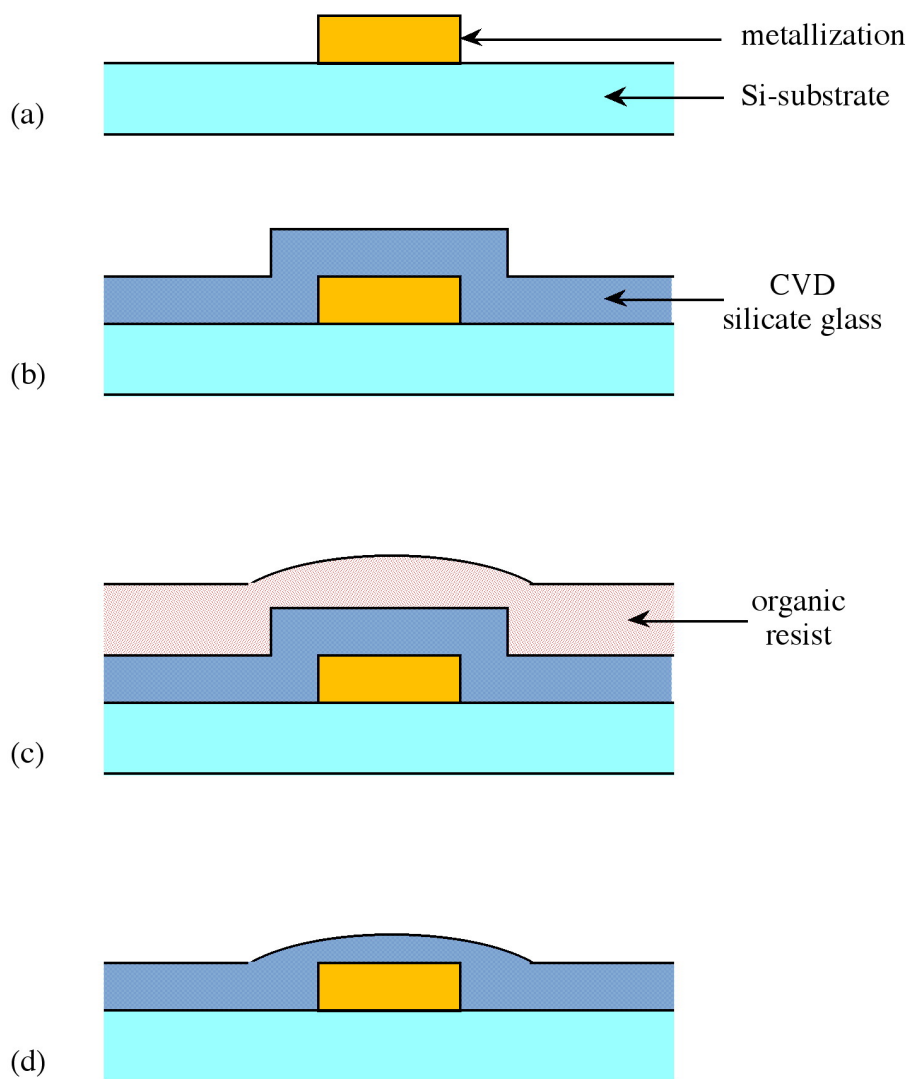
From the earliest days of solid state electronics it has been recognized that the presence or absence of surface states plays a decisive role in the usefulness of any semiconducting material. On the surface of any solid state material there are sites in which the coordination environment of the atoms is incomplete. These sites, commonly termed "dangling bonds", are the cause of the electronically active states which allow for the recombination of holes and electrons. This recombination occurs at energies below the bulk value, and interferes with the inherent properties of the semiconductor. In order to optimize the properties of a semiconductor device it is desirable to covalently satisfy all these surface bonds, thereby shifting the surface states out of the band gap and into the valence or conduction bands. Electronic passivation may therefore be described as a process which reduces the density of available electronic states present at the surface of a semiconductor, thereby limiting hole and electron recombination possibilities. In the case of silicon both the native oxide and other oxides admirably fulfill these requirements.

Chemical passivation requires a material that inhibits the diffusion of oxygen, water, or other species to the surface of the underlying semiconductor. In addition, the material is ideally hard and resistant to chemical attack. A perfect passivation material would satisfy both electronic and chemical passivation requirements.

Planarization

For the vast majority of electronic devices, the starting point is a substrate consisting of a flat single crystal wafer of semiconducting material. During processing, which includes the growth of both insulating and conducting films, the surface becomes increasingly non-planar. For example, a gate oxide in a typical MOSFET (see [\[link\]](#)) may be typically 100 - 250 Å thick, while the isolation or field oxide may be 10,000 Å. In order for the successful subsequent deposition of conducting layers (metallization) to occur without breaking metal lines (often due to the difficulty in maintaining step coverage), the surface must be flat and smooth. This process is called planarization, and can be carried out by a technique known as sacrificial etchback. The steps for this process are outlined in [\[link\]](#). An abrupt step ([\[link\]](#)a) is coated with a conformal layer of a low melting

dielectric, e.g., borophosphorosilicate glass, BPSG ([\[link\]](#)b), and subsequently a sacrificial organic resin ([\[link\]](#)c). The sample is then plasma etched such that the resin and dielectric are removed at the same rate. Since the plasma etch follows the contour of the organic resin, a smooth surface is left behind ([\[link\]](#)d). The planarization process thus reduces step height differentials significantly. In addition regions or valleys between individual metallization elements (vias) can be completely filled allowing for a route to producing uniformly flat surfaces, e.g., the BPSG film shown in [\[link\]](#).



Schematic representation of the planarization process.

A metallization feature (a) is CVD covered with silicate glass (b), and subsequently coated with an organic resin (c). After etching the resist a smooth silicate surface is produced (d).

The processes of planarization is vital for the development of multilevel structures in VLSI circuits. To minimize interconnection resistance and conserve chip area, multilevel metallization schemes are being developed in which the interconnects run in 3-dimensions.

Bibliography

- J. L. Vossen and W. Kern, *Phys. Today*, 1980, **33**, 26.
- S. K. Ghandhi, *VLSI Fabrication Principles, Silicon and Gallium Arsenide*, Wiley, Chichester, 2nd Ed. (1994).
- S. M. Sze, *Physics of Semiconductor Devices*, 2nd Edition, John Wiley & Sons, New York (1981).
- W. E. Beadle, J. C. C. Tsai, R. D. Plummer, *Quick Reference Manual for Silicon Integrated Circuit Technology*, Wiley, Chichester (1985).
- A. C. Adams and C. D. Capio, *J. Electrochem. Soc.*, 1981, **128**, 2630.

Oxidation of Silicon

Note: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Andrea Keys.

Introduction

In the fabrication of integrated circuits (ICs), the oxidation of silicon is essential, and the production of superior ICs requires an understanding of the oxidation process and the ability to form oxides of high quality. Silicon dioxide has several uses:

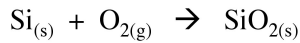
1. Serves as a mask against implant or diffusion of dopant into silicon.
2. Provides surface passivation.
3. Isolates one device from another (dielectric isolation).
4. Acts as a component in MOS structures.
5. Provides electrical isolation of multi-level metallization systems.

Methods for forming oxide layers on silicon have been developed, including thermal oxidation, wet anodization, chemical vapor deposition (CVD), and plasma anodization or oxidation. Generally, CVD is used when putting the oxide layer on top of a metal surface, and thermal oxidation is used when a low-charge density level is required for the interface between the oxide and the silicon surface.

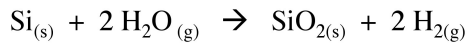
Oxidation of silicon

Silicon's surface has a high affinity for oxygen and thus an oxide layer rapidly forms upon exposure to the atmosphere. The chemical reactions which describe this formation are:

Equation:

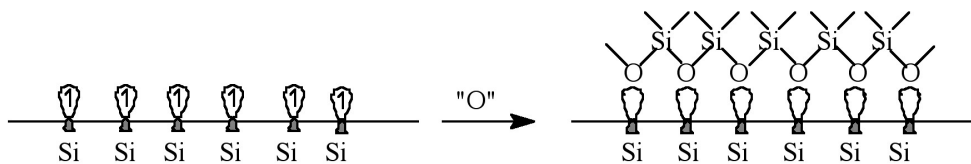


Equation:



In the first reaction a dry process is utilized involving oxygen gas as the oxygen source and the second reaction describes a wet process which uses steam. The dry process provides a "good" silicon dioxide but is slow and mostly used at the beginning of processing. The wet procedure is problematic in that the purity of the water used cannot be guaranteed to a suitable degree. This problem can be easily solved using a pyrogenic technique which combines hydrogen and oxygen gases to form water vapor of very high purity. Maintaining reagents of high quality is essential to the manufacturing of integrated circuits, and is a concern which plagues each step of this process.

The formation of the oxide layer involves shared valence electrons between silicon and oxygen, which allows the silicon surface to rid itself of "dangling" bonds, such as lone pairs and vacant orbitals, [\[link\]](#). These vacancies create mid-gap states between the valence and conduction bands, which prevents the desired band gap of the semiconductor. The Si-O bond strength is covalent (strong), and so can be used to achieve the loss of mid-gap states and passivate the surface of the silicon.



Removal of dangling bonds by oxidation of surface.

The oxidation of silicon occurs at the silicon-oxide interface and consists of four steps:

Diffusive transport of oxygen across the diffusion layer in the vapor phase adjacent to the silicon oxide-vapor interface.

Incorporation of oxygen at the outer surface into the silicon oxide film.

Diffusive transport across the silicon oxide film to its interface with the silicon lattice.

Reaction of oxygen with silicon at this inner interface.

As the Si-SiO₂ interface moves into the silicon its volume expands, and based upon the densities and molecular weights of Si and SiO₂, 0.44 Å Si is used to obtain 1.0 Å SiO₂.

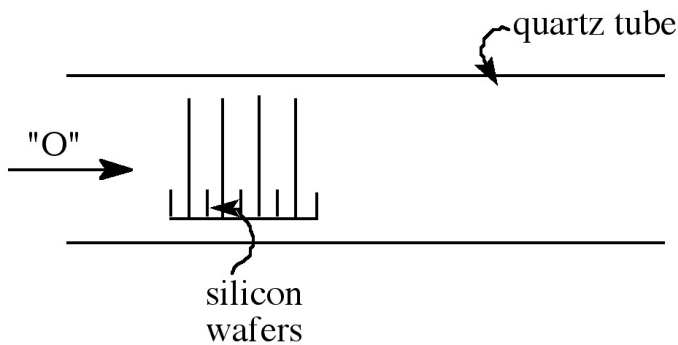
Pre-oxidation cleaning

The first step in oxidizing a surface of silicon is the removal of the native oxide which forms due to exposure to open air. This may seem redundant to remove an oxide only to put on another, but this is necessary since uncertainty exists as to the purity of the oxide which is present. The contamination of the native oxide by both organic and inorganic materials (arising from previous processing steps and handling) must be removed to prevent the degradation of the essential electrical characteristics of the device. A common procedure uses a H₂O-H₂O₂-NH₄OH mixture which removes the organics present, as well as some group I and II metals.

Removal of heavy metals can be achieved using a H₂O-H₂O₂-HCl mixture, which complexes with the ions which are formed. After removal of the native oxide, the desired oxide can be grown. This growth is useful because it provides: chemical protection, conditions suitable for lithography, and passivation. The protection prevents unwanted reactions from occurring and the passivation fills vacancies of bonds on the surface not present within the interior of the crystal. Thus the oxidation of the surface of silicon fulfills several functions in one step.

Thermal oxidation

The growth of oxides on a silicon surface can be a particularly tedious process, since the growth must be uniform and pure. The thickness wanted usually falls in the range 50 - 500 Å, which can take a long time and must be done on a large scale. This is done by stacking the silicon wafers in a horizontal quartz tube while the oxygen source flows over the wafers, which are situated vertically in a slotted paddle (boat), see [\[link\]](#). This procedure is performed at 1 atm pressure, and the temperature ranges from 700 to 1200 °C, being held to within ± 1 °C to ensure uniformity. The choice of oxidation technique depends on the thickness and oxide properties required. Oxides that are relatively thin and those that require low charge at the interface are typically grown in dry oxygen. When thick oxides are required (> 0.5 mm) are desired, steam is the source of choice. Steam can be used at wide range of pressures (1 atm to 25 atm), and the higher pressures allow thick oxide growth to be achieved at moderate temperatures in reasonable amounts of time.



Horizontal diffusion tube showing the oxidation of silicon wafers at 1 atm pressure.

The thickness of SiO₂ layers on a Si substrate is readily determined by the color of the film. [\[link\]](#) provides a guideline for thermal grown oxides.

Film thickness (μm)	Color	Film thickness (μm)	Color
0.05	tan	0.63	violet-red
0.07	brown	0.68	"bluish"
0.10	dark violet to red-violet	0.72	blue-green to gree
0.12	royal blue	0.77	"yellowish"
0.15	light blue to metallic blue	0.80	orange
0.17	metallic to light yellow-green	0.82	salmon
0.20	light gold	0.85	light red-violet
0.22	gold	0.86	violet
0.25	orange to melon	0.87	blue violet
0.27	red-violet	0.89	blue
0.30	blue to violet blue	0.92	blue-green
0.31	blue	0.95	yellow-green
0.32	blue to blue-green	0.97	yellow
0.34	light green	0.99	orange

0.35	green to yellow-green	1.00	carnation pink
0.36	yellow-green	1.02	violet red
0.37	green-yellow	1.05	red-violet
0.39	yellow	1.06	violet
0.41	light orange	1.07	blue-violet
0.42	carnation pink	1.10	green
0.44	violet-red	1.11	yellow-green
0.46	red-violet	1.12	green
0.47	violet	1.18	violet
0.48	blue-violet	1.19	red-violet
0.49	blue	1.21	violet-red
0.50	blue green	1.24	carnation pink to salmon
0.52	green	1.25	orange
0.54	yellow-green	1.28	"yellowish"
0.56	green-yellow	1.32	sky blue to green-blue
0.57	"yellowish"	1.40	orange

0.58	light orange to pink	1.46	blue-violet
0.60	carnation pink	1.50	blue

Color chart for thermally grown SiO₂ films observed under daylight fluorescent lighting.

High pressure oxidation

High pressure oxidation is another method of oxidizing the silicon surface which controls the rate of oxidation. This is possible because the rate is proportional to the concentration of the oxide, which in turn is proportional to the partial pressure of the oxidizing species, according to Henry's law, [\[link\]](#), where C is the equilibrium concentration of the oxide, H is Henry's law constant, and p_O is the partial pressure of the oxidizing species.

Equation:

$$C = H_{(pO)}$$

This approach is fast, with a rate of oxidation ranging from 100 to 1000 mm/h, and also occurs at a relatively low temperature. It is a useful process, preventing dopants from being displaced and also forms a low number of defects, which is most useful at the end of processing.

Plasma oxidation

Plasma oxidation and anodization of silicon is readily accomplished by the use of activated oxygen as the oxidizing species. The highly reactive oxygen is formed within an electrical discharge or plasma. The oxidation is carried out in a low pressure (0.05 - 0.5 Torr) chamber, and the plasma is produced either by a DC electron source or a high-frequency discharge. In simple plasma oxidation the sample (i.e., the silicon wafer) is held at

ground potential. In contrast, anodization systems usually have a DC bias between the sample and an electrode with the sample biased positively with respect to the cathode. Platinum electrodes are commonly used as the cathodes.

There have been at least 34 different reactions reported to occur in an oxygen plasma, however, the vast majority of these are inconsequential with respect to the formation of active species. Furthermore, many of the potentially active species are sufficiently short lived that it is unlikely that they make a significant contribution. The primary active species within the oxygen plasma are undoubtedly O^- and O^{2+} . Both being produced in near equal quantities, although only the former is relevant to plasma anodization. While these species may be active with respect to surface oxidation, it is more likely that an electron transfer occurs from the semiconductor surface yields activated oxygen species, which are the actual reactants in the oxidation of the silicon.

The significant advantage of plasma processes is that while the electron temperature of the ionized oxygen gas is in excess of 10,000 K, the thermal temperatures required are significantly lower than required for the high pressure method, i.e., $< 600^\circ\text{C}$. The advantages of the lower reaction temperatures include: the minimization of dopant diffusion and the impediment of the generation of defects. Despite these advantages there are two primary disadvantages of any plasma based process. First, the high electric fields present during the processes cause damage to the resultant oxide, in particular, a high density of interface traps often result. However, post annealing may improve film quality. Second, the growth rates of plasma oxidation are low, typically 1000 \AA/h . This growth rate is increased by about a factor of 10 for plasma anodization, and further improvements are observed if 1 - 3% chlorine is added to the oxygen source.

Masking

A selective mask against the diffusion of dopant atoms at high temperatures can be found in a silicon dioxide layer, which can prove to be very useful in integrated circuit processing. A predeposition of dopant by ion

implantation, chemical diffusion, or spin-on techniques typically results in a dopant source at or near the surface of the oxide. During the initial high-temperature step, diffusion in the oxide must be slow enough with respect to diffusion in the silicon that the dopants do not diffuse through the oxide in the masked region and reach the silicon surface. The required thickness may be determined by experimentally measuring, at a particular temperature and time, the oxide thickness necessary to prevent the inversion of a lightly doped silicon substrate of opposite conductivity. To this is then added a safety factor, with typical total values ranging from 0.5 to 0.7 μm . The impurity masking properties result when the oxide is partially converted into a silica impurity oxide "glass" phase, and prevents the impurities from reaching the SiO_2 -Si interface.

Bibliography

- M. M. Atalla, in *Properties of Elemental and Compound Semiconductors*, Ed. H. Gatos, Interscience: New York (1960).
- S. K. Ghandhi, *VLSI Fabrication Principles, Silicon and Gallium Arsenide*, Wiley, Chichester, 2nd Ed. (1994).
- S. M. Sze, *Physics of Semiconductor Devices*, 2nd Edition, John Wiley & Sons, New York (1981).
- D. L. Lile, *Solid State Electron.*, 1978, **21**, 1199.
- W. E. Spicer, P. W. Chye, P. R. Skeath, and C. Y. Su, I. Lindau, *J. Vac. Sci. Technol.*, 1979, **16**, 1422.
- V. Q. Ho and T. Sugano, *IEEE Trans. Electron Devices*, 1980, **ED-27**, 1436.
- J. R. Hollanhan and A. T. Bells, *Techniques and Applications of Plasma Chemistry*, Wiley, New York (1974).
- R. P. H. Chang and A. K. Sinha, *Appl. Phys. Lett.*, 1976, **29**, 56.

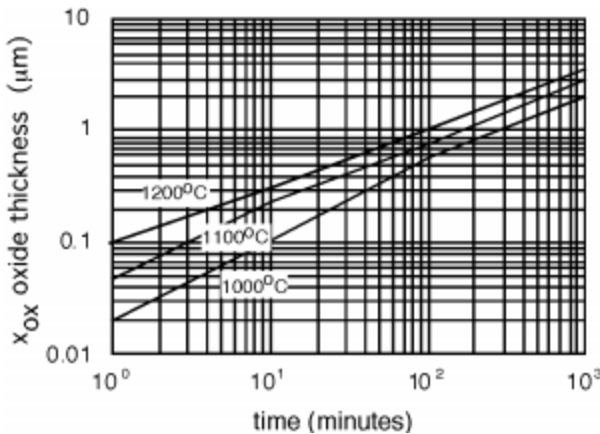
Photolithography

Note: This module is based upon the Connexions module entitled *Photolithography* by Bill Wilson.

Actually, implants (especially for moats) are usually done at a sufficiently high energy so that the dopant (phosphorus) is already pretty far into the substrate (often several microns or so), even before the diffusion starts. The anneal/diffusion moves the impurities into the wafer a bit more, and as we shall see also makes the n-region grow larger.

"The n-region"! We have not said a thing about how we make our moat in only certain areas of the wafer. From the description we have so far, it seems we have simply built an n-type layer over the whole surface of the wafer. This would be bad! We need to come up with some kind of "window" to only permit the implanting impurities to enter the silicon wafer where we want them and not elsewhere. We will do this by constructing an implantation "barrier".

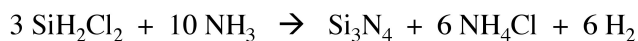
To do this, the first thing we do is grow a layer of silicon dioxide over the entire surface of the wafer. We talked about oxide growth when we were discussing MOSFETs but let's go into a little more detail. You can grow oxide in either a dry oxygen atmosphere, or in an atmosphere which contains water vapor, or steam. In [\[link\]](#), we show oxide thickness as a function of time for growth with steam. Dry O₂ does not behave too much differently, the rate is just somewhat slower.



A plot of oxide thickness as a function of time.

On top of the oxide, we are now going to deposit yet another material. This is silicon nitride, Si_3N_4 or just plain "nitride" as it is usually called. Silicon nitride is deposited through a method called chemical vapor deposition or "CVD". The usual technique is to react dichlorosilane and ammonia in a hot walled low pressure chemical vapor deposition system (LPCVD). The reaction is:

Equation:



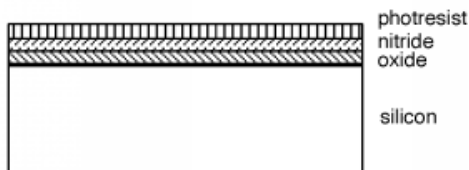
Silicon nitride is a good barrier for impurities, oxygen and other things which do not want to get into the wafer. Take a look at [\[link\]](#) and see what we have so far. A word about scale and dimensions. The silicon wafer is about 250 μm thick (about 0.01") since it has to be strong enough not to break as it is being handled. The two deposited layers are each about 1 μm thick, so they should actually be drawn as lines thinner than the other lines in the figure. This would obviously make the whole idea of a sketch ridiculous, so we will leave things distorted as they are, keeping in mind that the deposited and diffused layers are actually much thinner than the rest

of wafer, which really does not do anything except support the active circuits up on top.



Initial wafer configuration.

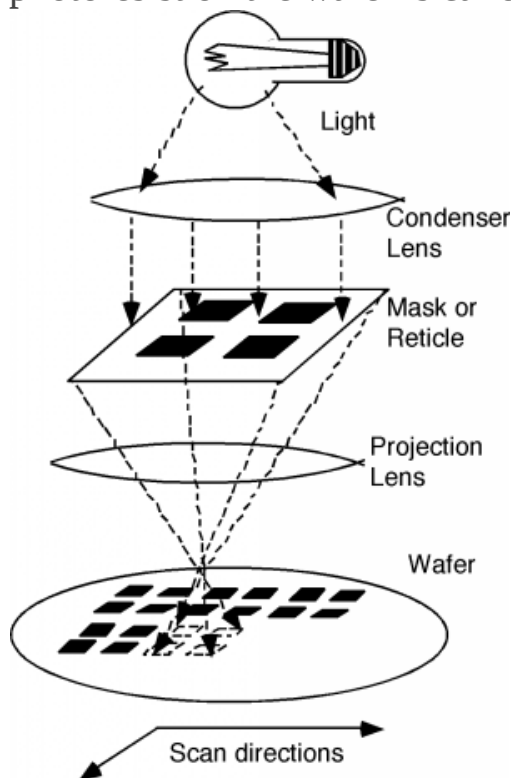
Now what we want to do is remove part of the nitride, so we can make our n-well, but not put in phosphorous where do not want it. We do this with a processes called *photolithography* and *etching* respectively. First thing we do is coat the wafer with yet another layer of material. This is a liquid called photoresist and it is applied through a process called spin-coating. The wafer is put on a vacuum chuck, and a layer of liquid photoresist is sprayed uncap of the wafer. The chuck is then spun rapidly, getting to several thousand RPM in a small fraction of a second. Centrifugal force causes the resist to spread out uniformly across the wafer surface. The solvent for the photoresist is quite volatile and so the layer of photoresist dries while the wafer is still spinning, resulting in a thin, uniform coating across the wafer [\[link\]](#).



After the photoresist is spun on.

The name "photoresist" gives some clue as to what this stuff is. Basically, photoresist is a polymer mixed with some kind of light sensitizing compound. In positive photoresist, wherever light strikes it, the polymer is weakened, and it can be more easily removed with a solvent during the development process. Conversely, negative photoresist is strengthened when it is illuminated with light, and is more resistant to the solvent than is the unilluminated photoresist. Positive resist is so-called because the image of the developed photoresist on the wafer looks just like the mask that was used to create it. Negative photoresist makes an image which is the opposite of what the mask looks like.

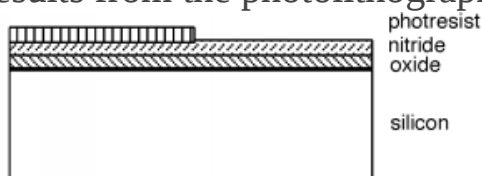
We have to come up with some way of selectively illuminating certain portions of the photoresist. Anyone who has ever seen a projector know how we can do this. But, since we want to make small things, not big ones, we will change around our projector so that it makes a smaller image, instead of a bigger one. The instrument that projects the light onto the photoresist on the wafer is called a projection printer or stepper [\[link\]](#).



A schematic of a stepper configuration.

As shown in [\[link\]](#), the stepper consists of several parts. There is a light source (usually a mercury vapor lamp, although ultra-violet excimer lasers are also starting to come into use), a condenser lens to image the light source on the mask or reticle. The mask contains an image of the pattern we are trying to place on the wafer. The projection lens then makes a reduced (usually 5x) image of the mask on the wafer. Because it would be far too costly, if not just plain impossible, to project onto the whole wafer all at once, only a small selected area is printed at one time. Then the wafer is scanned or stepped into a new position, and the image is printed again. If previous patterns have already been formed on the wafer, TV cameras, with artificial intelligence algorithms are used to align the current image with the previously formed features. The stepper moves the whole surface of the wafer under the lens, until the wafer is completely covered with the desired pattern. A stepper is one of the most important pieces of equipment in the whole IC fab however, since it determines what the minimum feature size on the circuit will be.

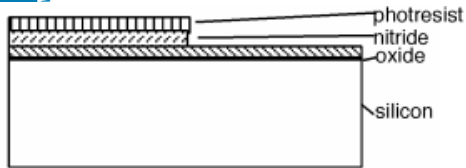
After exposure, the photoresist is placed in a suitable solvent, and "developed". Suppose for our example the structure shown in [\[link\]](#) is what results from the photolithographic step.



After photoresist
exposure and
development.

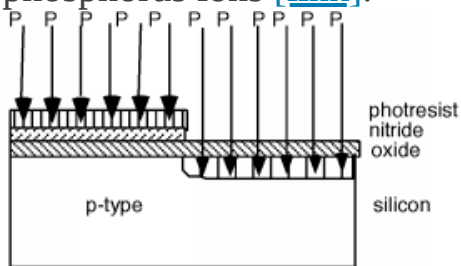
The pattern that was used in the photolithographic (PL) step exposed half of our area to light, and so the photoresist (PR) in that region was removed upon development. The wafer is now immersed in a hydrofluoric acid (HF)

solution. HF acid etches silicon nitride quite rapidly, but does not etch silicon dioxide nearly as fast, so after the etch we have what we see in [\[link\]](#).



After the nitride etch step.

We now take our wafer, put it in the ion implanter and subject it to a "blast" of phosphorus ions [\[link\]](#).



Implanting phosphorus.

The ions go right through the oxide layer on the RHS, but stick in the resist/nitride layer on the LHS of our structure.

Optical Issues in Photolithography

Note: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Zane Ball.

Introduction

Photolithography is one of the most important technology in the production of advanced integrated circuits. It is through photolithography that semiconductor surfaces are patterned and the circuits formed. In order to make extremely small features, on the order of the wavelength of the light, advanced optical techniques are used to transfer a pattern from a mask onto the surface. A polymeric film or *resist*, is modified by the light and records the information in a process not dissimilar to ordinary photography.

An illustration of the photolithographic process is shown in [\[link\]](#). The process follows the following basic steps:

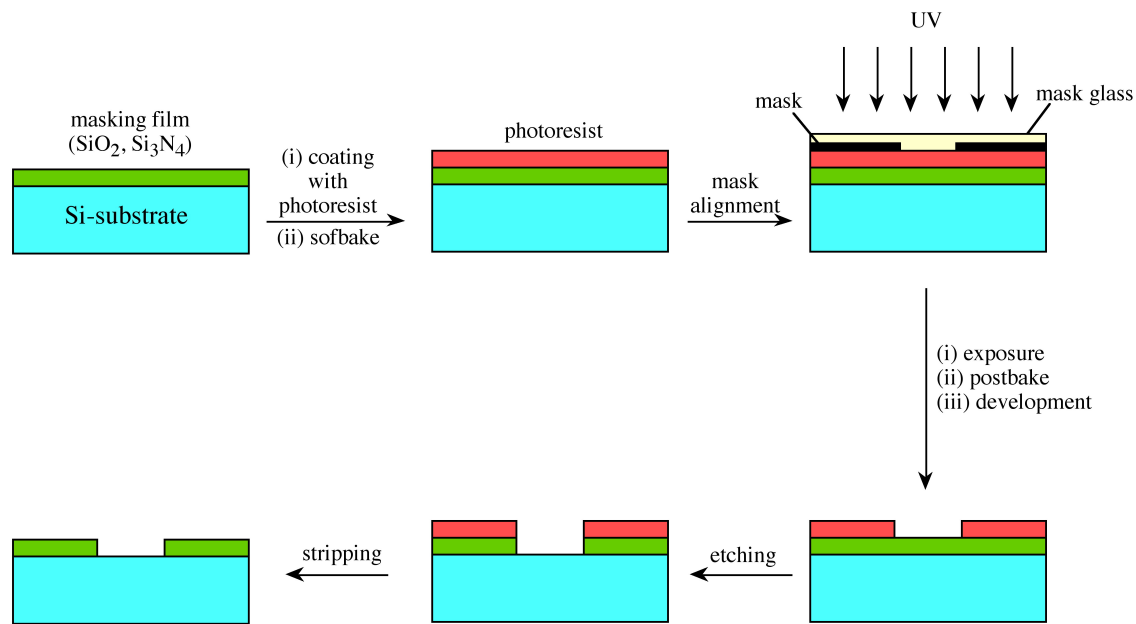
The wafer is spin coated with resist to form a uniform $\sim 1\text{ }\mu\text{m}$ thin film of resist on the surface.

The wafer is exposed with ultraviolet light through a mask which contains the desired pattern. In the simplest processes the mask is simply placed over the wafer, but advanced sub-micron technologies require the pattern to be imaged through a complex optical system.

The photoresist is developed and the irradiated area is washed away (positive resist) or the unirradiated area is washed away (negative resist).

Processing (etching, deposition etc.)

Remaining resist is stripped.



Steps in optical printing using photolithography.

In addition to being possibly the most important semiconductor process step, photolithography is also the most expensive technology in semiconductor manufacturing. This expense is the result of two considerations:

1. The optics in photolithography tools are expensive where a single lens can cost a \$1 million or more
2. Each chip (often referred to as a "die") must be exposed individually unlike other semiconductor processes such as CVD where an entire wafer can be processed at a time or oxidation processes where many wafers can be processed simultaneously.

This means that not only are photolithography machines the most expensive of semiconductor processing equipment, but more of them are needed in order to maintain throughput.

Optical issues in photolithography

The critical dimension and depth of focus

A semiconductor process technology is often described by a characteristic length known as the critical dimension. The critical dimension (CD) is the smallest feature that needs to be patterned on the surface. The exact definition varies from process to process but is often the channel length of the smallest transistor (typical of a memory chip) or the width of the smallest metal interconnection line (logic chips). This critical dimension is defined by the photolithographic process and is perhaps the most important figure of merit in the manufacture of integrated circuits. Making the critical dimension smaller is the primary focus of improving semiconductor technology for the following reasons:

1. Making the CD smaller dramatically increases the number of devices per unit area and this increase goes with the square of the CD (i.e., a reduction in CD by a factor of 2 generates 4 times the number of devices).
2. Making the CD smaller of a device already in production will make a smaller chip. This means that the number of chips per wafer increases dramatically, and since costs generally scale with the number of wafers and not the number of chips to a wafer, costs are dramatically reduced.
3. Smaller devices are faster.

Therefore, improvements in lithography technology translate directly into better, faster, more complex circuits at lower cost.

Having established the importance of the critical dimension it is important to understand what features of a photolithography system impact. The theory behind projection lithography is very well known, dating from the original analysis of the microscope by Abbe. It is, in fact, the Abbe sine condition that dictates the critical dimension:

Equation:

$$CD_{Coherent} = 0.82 \frac{\lambda}{n \sin(\theta)}$$

$$CD_{Incoherent} = 0.61 \frac{\lambda}{n \sin(\theta)}$$

where the two expressions refer to the limit of a purely coherent illuminating source and purely incoherent source respectively, and λ is the vacuum wavelength of the illuminating light source, n the index of refraction of the objective lens, and θ refers to the angle between the axis of the lens and the line from the back focal point to the aperture of the entrance of the lens. The quantity in the denominator, $n \sin(\theta)$ is referred to as the numerical aperture or NA. As the degree of coherence can be adjusted in a lithography system, the critical dimension is usually written more generally as:

Equation:

$$CD = k_1 \frac{\lambda}{n \sin(\theta)}$$

From this equation, we begin to see what can be done to reduce the critical dimension of a lithography system:

1. Change the wavelength of the source.
2. Increase the numerical aperture (NA).
3. Reduce k_1 .

Before we discuss how this is accomplished, we must consider one other key quantity, the depth of focus or DOF. The depth of focus is the length along the axis in which a sharp image exists. Naturally a large DOF is desirable for ease of alignment, since the entire dye must with lie within this region. In reality, however, the more meaningful constraint is that the DOF must be thicker than the resist layer so that the entire volume of resist is exposed and can be developed. Also, if the surface morphology of the device dictates that the resist to be exposed is not planar, then the DOF

must be large enough so that all features are properly illuminated. Current resists must be 1 μm in thickness in order to have the necessary etch resistance, so this can be considered a minimum value for an acceptable DOF. The depth of focus can also be expressed as a function of numerical aperture and wavelength:

Equation:

$$DOF = k_2 \frac{\lambda}{[n \sin(\theta)]^2}$$

If we desire to minimize the critical dimension simply by making optics of large numerical aperture that we will simultaneously reduce the depth of focus and at a much faster rate owing to the dependence on the square of the numerical aperture.

These two quantities, DOF and CD, provide the direction in lithography and semiconductor processing as a whole. For example, a design with an improved surface planarity or a new resist that is effective at smaller thicknesses would allow for a smaller depth of focus which would in turn allow for a larger numerical aperture implying a smaller critical dimension. The resist, the source wavelength, and the optical delivery system all affect the critical dimension and that further refinements require a multifaceted approach to improving lithography systems. What also must be realized is that, as far as the optical system is concerned, virtually all that can be done with conventional optics has been done and that fundamental restraints on k_1 have been reached.

Wavefront engineering

One way to get around the fundamental limitations of an imaging system illustrated in [\[link\]](#) is through one of a variety of techniques often termed *wavefront engineering*. Here, not only is the amplitude mapped from the object plane to the image plane, but the phase structure of the light going through the mask is manipulated to improve the contrast and allow for effective values of k_1 lower than the theoretical minimum for uniform

illumination. The most important example of these techniques is the phase shift mask or PSM. Here the mask consists of two types of areas, those that allow light to pass through unaffected and some regions where the amplitude of the light is unaffected but its phase is shifted. The resulting electric fields will then sum to zero in some places where use of an ordinary mask would have resulted in a positive intensity.

There are many problems with the practical introduction of various phase shifting techniques. Construction of masks with phase shifting elements (usually a thin layer of PMMA) is difficult and expensive. Mask damage, already a key problem in conventional production techniques, becomes an even greater issue as traditional mask repair techniques can no longer be used. Also identifying errors in a mask is made more difficult by the odd design.

Interaction with resists

The ultimate resolution of a photolithographic process is not dependent on optics alone, but also on the interaction with the resist. One of the key concerns, particularly as wavelengths of sources become shorter, is the ability of the source light to penetrate the resist film. Many polymers absorb strongly in the UV which can limit the interaction to the surface. In such a case only a thin layer of the polymer is exposed and the pattern may not be fully uncovered during developing. One important property of resist is the presence of *saturable absorption*. Saturable absorbers are those absorption sites in the polymer that when excited to a higher state remain there for relatively long periods of time and do not continue to absorb into higher states. If only saturable absorption is present in a polymer film, then continued irradiation eventually leads to transparency as all absorption sites will be saturated. This allows light penetration through the resist film with full exposure to the substrate surface.

Full penetration of the film leads to a second problem, multiple reflection interference. This occurs when light which has penetrated the film to the substrate is then reflected back towards the surface. The result is a standing wave interference pattern which causes uneven exposure through the film.

The problem becomes more severe as optical limits are approached where feature size is approximately equal to the wavelength of the light source meaning such standing waves are the same size as the irradiated features. In the most advanced lithography techniques such as 248 nm lithography with excimer lasers, a special anti-reflectance coating must be laid down before the resist is deposited. Development of an AR coating that has no adverse effects during the exposure and development process is difficult.

One completely new approach to photolithography resists are top-surface-imaged resists or TSI resists. These processes do not require light penetration through the whole volume of resist. In a TSI resist, a silyl amine is selectively in-diffused from the gas phase into a phenolic polymer in response to the laser irradiation. This diffusion process creates a silyl ether, and development takes place in the form of an oxygen plasma etch, sometimes termed 'dry developing'. Depth of focus limitations are thus avoided as exposure is necessary only at the surface of the resist layer, and the resolution of the etching process determines the final resist profile. Such a technique has tremendous advantages, particularly as source wavelengths become shorter and transparent polymers more rare. Such a resist has a clear optical advantage as well since the image need only be formed at the surface of the resist layer reducing the DOF needed to 100 nm or less, allowing for larger numerical aperture lithography systems with smaller critical dimensions.

Light sources

Current photolithography techniques in production utilize ultraviolet lamps as the light source. In the most advanced production facilities, 0.35 μm mercury i-line technology is used. For the next generation of chips such as 64 Mbit DRAMS better performance is necessary and either i-line technology combined with PSM or a new light source is required. Certainly for the 256 Mbit generation using 0.25 μm technology, the i-line source is no longer adequate. The apparent successor is the 248 nm KrF laser, which entered the most advanced production facilities in the late 1990s. KrF technology is often referred to in the literature as Deep UV or DUV lithography. For further shrinkage to 0.18 μm technology, the ArF excimer

laser at 193 nm will likely be used with the transition likely to take place in the first few years of the next decade.

At critical dimensions lower than 0.18 - 0.1 μm and below, a whole host of technological problems will need to be overcome in every stage of manufacturing including photolithography. One likely scheme for future lithography is to use X-rays where the wavelength of the light is so much smaller than the feature size such that proximity printing can be used. This is where the mask is placed close to the surface and an X-ray source is scanned across using no optics. Common X-ray sources for such techniques include synchrotron radiation and laser produced plasmas. It has also been widely suggested that the cost of implementing X-ray or other post-optical techniques together with the increased cost of every other manufacturing process step will make improvements beyond 0.1 μm cost prohibitive where benefits in increased circuit speed and density will be dwarfed by massive manufacturing cost. It is noted however that such predictions have been made in the past with regard to other technological barriers.

Bibliography

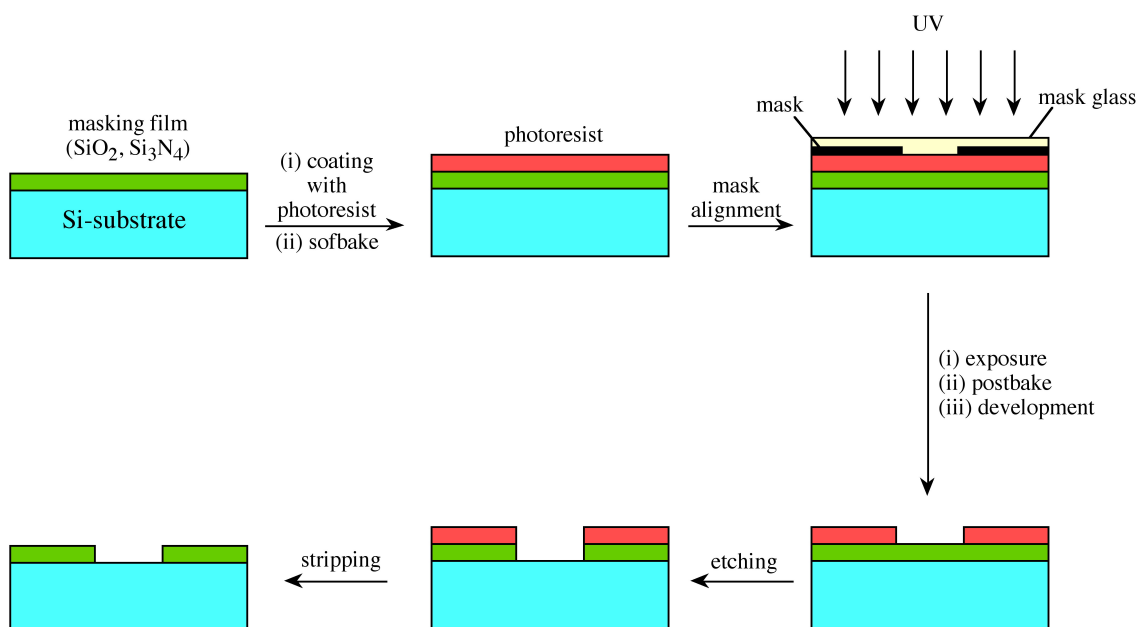
- M. Born and E. Wolf, *Principles of Optics 6th Edition*, Pergamon Press, New York (1980).
- M. Nakase, *IEICE Trans. Electron.*, 1993, **E76-C**, 26.
- M. Rothschild, A. R. Forte, M. W. Horn, R. R. Kunz, S. C. Palmateer, and J. H. C. Sedlacek, *IEEE J. Selected Topics in Quantum Electronics*, 1995, **1**, 916.

Composition and Photochemical Mechanisms of Photoresists

Note: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Angela Cindy Wei.

Photolithography

In photolithography, a pattern may be transferred onto a photoresist film by exposing the photoresist to light through a mask of the pattern. In the semiconductor industry, the photolithographic procedure includes the following steps as illustrated in [\[link\]](#): coating a base material with photoresist, exposing the resist through a mask to light, developing the resist, etching the exposed areas of the base, and stripping the remaining resist off.



Steps in optical printing using photolithography.

Upon exposure to light, the photoresist may become more or less soluble depending on the chemical properties of the particular resist material. The photochemical reactions include chain scission, cross-linking, and the rearrangement of molecules. If the exposed areas of the photoresist become more soluble, then it is a positive resist; conversely, if the exposed resist becomes less soluble, then it is a negative resist. In developing the photoresist, the more soluble material is removed leaving a positive or a negative image of the mask pattern.

Photoresist

Photoresists were initially developed for the printing industry. In the 1920s, the application of photoresists spread to the printed circuit board industry. Photoresists for semiconductor use were first developed in the 1950s; Kodak developed commercial negative photoresists and shortly after, Shipley developed a line of positive resists. Several other companies have entered the market since that time in hopes of manufacturing resist products which meet the increasing demands of the semiconductor industry: narrower line widths, fewer defects, and higher production rates.

Photoresist composition

Several functional requirements must be met for a photoresist to be used in the semiconductor industry. Photoresist polymers must be soluble for easy deposition onto a substrate by spin-coating. Good photoresist-substrate adhesion properties are required to minimize undercutting, to maintain edge acuity, and to control the feature sizes. The photoresist must be chemically resistant to whichever etchants are to be used. Sensitivity of the photoresist to a particular light source is essential to the functionality of a photoresist. The speed at which chemical changes occur in a photoresist is its contrast. The contrast of a resist is dependent on the molecular weight distribution of

the polymers: a broad molecular weight distribution results in a low contrast resist. High contrast resists produce higher resolution images.

The four basic components of a photoresist are the polymer, the solvent, sensitizers, and other additives. The role of the polymer is to either polymerize or photosolubilize when exposed to light. Solvents allow the photoresist to be applied by spin-coating. The sensitizers control the photochemical reactions and additives may be used to facilitate processing or to enhance material properties. Photochemical changes to polymers are essential to the functionality of a photoresist. Polymers are composed primarily of carbon, hydrogen, and oxygen-based molecules arranged in a repeated pattern. Negative photoresists are based on polyisoprene polymers; negative resist polymers are not chemically bonded to each other, but upon exposure to light, the polymers crosslink, or polymerize. Positive photoresists are formulated from phenol-formaldehyde novolak resins; the positive resist polymers are relatively insoluble, but upon exposure to light, the polymers undergo photosolubilization.

Solvents are required to make the photoresist a liquid, which allows the resist to be spun onto a substrate. The solvents used in negative photoresists are non-polar organic solvents such as toluene, xylene, and halogenated aliphatic hydrocarbons. In positive resists, a variety of organic solvents such as ethyl cellosolve acetate, ethoxyethyl acetate, diglyme, or cyclohexanone may be used.

Photosensitizers are used to control or cause polymer reactions resulting in the photosolubilization or crosslinking of the polymer. The sensitizers may also be used to broaden or narrow the wavelength response of the photoresist. Bisazide sensitizers are used in negative photoresists while positive photoresists utilize diazonaphthoquinones. One measure of photosensitizers is their quantum efficiencies, the fraction of photons which result in photochemical reactions; the quantum efficiency of positive diazonaphthoquinone photoresist sensitizers has been measured to be 0.2 - 0.3 and the quantum efficiency of negative bis-arylazide sensitizers is in the range of 0.5 - 1.0.

Additives are also introduced into photoresists depending on the specific needs of the application. Additives may be used to increase photon

absorption or to control light within the resist film. Adhesion promoters such as hexamethyldisilazane and additives to improve substrate coating are also commonly used.

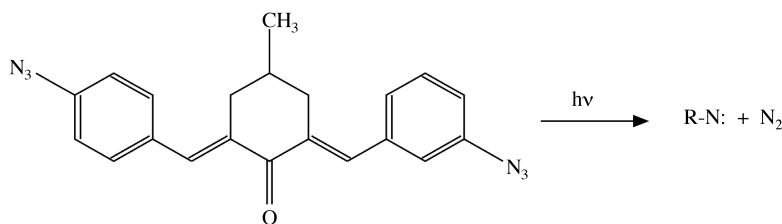
Negative photoresist chemistry

The matrix resin material used in the formulation of these (negative) resists is a synthetic rubber obtained by a Ziegler-Natta polymerization of isoprene which results in the formation of poly(cis-isoprene). Acid-catalyzation of poly(cis-isoprene) produces a partially cyclized polymer material; the cyclized polymer has a higher glass transition temperature, better structural properties, and higher density. On the average, microelectronic resist polyisoprenes contain 1-3 rings per cyclic unit, with 5-20% unreacted isoprene units remaining'. The resultant material is extremely soluble in non-polar, organic solvents including toluene, xylene, and halogenated aliphatic hydrocarbons.

The condensation of para-azidobenzaldehyde with a substituted cyclohexanone produces bis-arylazide sensitizers. To maximize the absorption of a particular light source, the absorbance spectrum of the photoresist may be shifted by making structural modifications to the sensitizers; for example, by using substituted benzaldehydes, the absorption peak may be shifted to longer wavelengths. A typical bisazide-cyclized polyisoprene photoresist formulation may contain 97 parts cyclized polyisoprene to 3 parts bisazide in a (10 wt%) xylene solvent.

All negative photoresists function by cross-linking a chemically reactive polymer via a photosensitive agent that initiates the chemical cross-linking reaction. In the bisazide-cyclized polyisoprene resists, the absorption of photons by the photosensitive bisazide in the photoresist results in an insoluble crosslinked polymer. Upon exposure to light, the bisazide sensitizers decompose into nitrogen and highly reactive chemical intermediates, called nitrenes [\[link\]](#). The nitrenes react to produce polymer linkages and three-dimensional cross-linked structures that are less soluble in the developer solution.

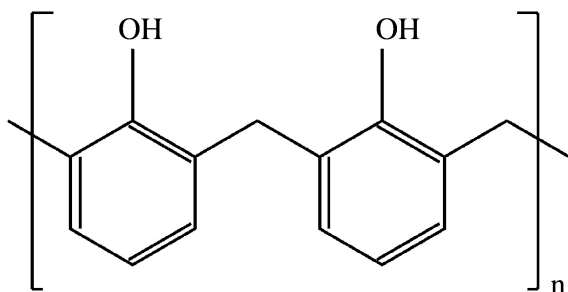
Equation:



Positive photoresist chemistry

Positive photoresist materials originally developed for the printing industry have found use in the semiconductor industry. The commonly used novolac resins (phenol-formaldehyde copolymer) and (photosensitive) diazoquinone both were products of the printing industry.

The novolac resin is a copolymer of a phenol and formaldehyde ([\[link\]](#)). Novolac resins are soluble in common organic solvents (including ethyl cellosolve acetate and diglyme) and aqueous base solutions. Commercial resists usually contain meta-cresol resins formed by the acid-catalyzed condensation of meta-cresol and formaldehyde.



Structure of a novolac resin.

The positive photoresist sensitizers are substituted diazonaphthoquinones. The choice of substituents affects the solubility and the absorption

characteristics of the sensitizers. Common substituents are aryl sulfonates. The diazoquinones are formed by a reaction of diazonaphthoquinone sulfonyl chloride with an alcohol to form sulfonate ester; the sensitizers are then incorporated into the resist via a carrier or bonded to the resin. The sensitizer acts as a dissolution inhibitor for the novolac resin and is base-insoluble. The positive photoresist is formulated from a novolac resin, a diazonaphthoquinone sensitizer, and additives dissolved in a 20 - 40 wt% organic solvent. In a typical resist, up to 40 wt% of the resist may be the sensitizer.

The photochemical reaction of quinonediazide is illustrated in [\[link\]](#). Upon absorption of a photon, the quinonediazide decomposes through Wolff rearrangement, specifically a *S₂* reaction, and produces gaseous nitrogen as a by-product. In the presence of water, the decomposition product forms an indene carboxylic acid, which is base-soluble. However, the formation of acid may not be the reason for increased solubility; the release of nitrogen gas produces a porous structure through which the developer may readily diffuse, resulting in increased solubility.

Equation:

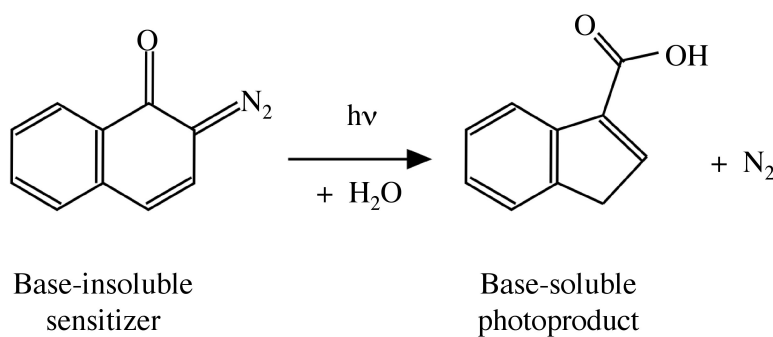


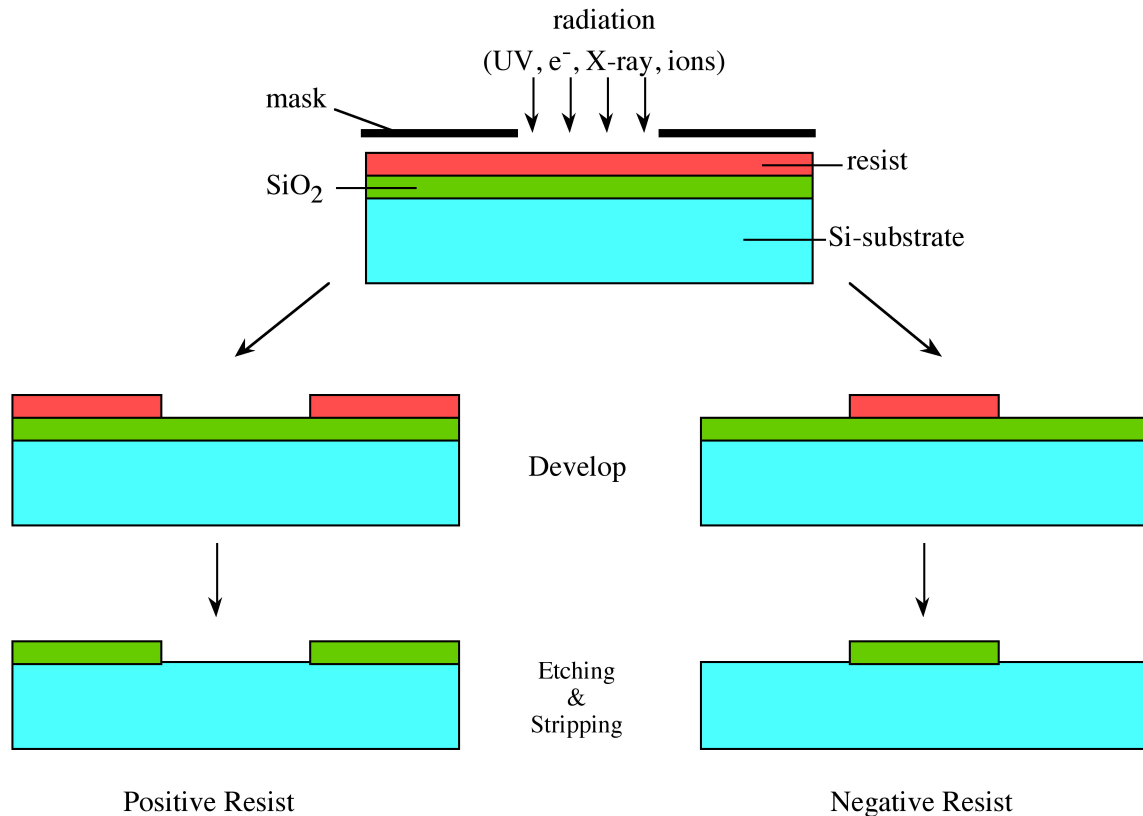
Image reversal

By introducing an additive to the novolac resins with diazonaphthaquinones sensitizers, the resultant photoresist may be used to form a negative image. A small amount of a basic additive such as monazoline, imidazole, and triethylamine is mixed into a positive novolac resist. Upon exposure to

light, the diazonaphthaquiones sensitizer forms an indene carboxylic acid. During the subsequent baking process, the base catalyzes a thermal decarboxylation, resulting in a substituted indene that is insoluble in aqueous base. Then, the resist is flood exposed destroying the dissolution inhibitors remaining in the previously unexposed regions of the resist. The development of the photoresist in aqueous base results in a negative image of the mask.

Comparison of positive and negative photoresists

Into the 1970s, negative photoresist processes dominated. The poor adhesion and the high cost of positive photoresists prevented its widespread use at the time. As device dimensions grew smaller, the advantages of positive photoresists, better resolution and pinhole protection, suited the changing demands of the semiconductor industry and in the 1980s the positive photoresists came into prominence. A comparison of negative and positive photoresists is given in [\[link\]](#).



A comparison of negative and positive photoresists.

The better resolution of positive resists over negative resists may be attributed to the swelling and image distortion of negative resists during development; this prevents the formation of sharp vertical walls of negative resist. Disadvantages of positive photoresists include a higher cost and lower sensitivity.

Positive photoresists have become the industry choice over negative photoresists. Negative photoresists have much poorer resolution and the positive photoresists exhibit better etch resistance and better thermal stability. As optical masking processes are still preferred in the semiconductor industry, efforts to improve the processes are ongoing. Currently, researchers are studying various forms of chemical amplification to increase the photon absorption of photoresists.

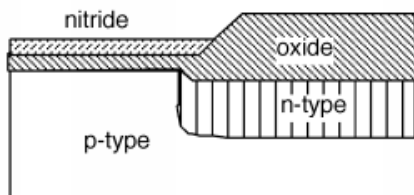
Bibliography

- W.M. Alvino, *Plastics For Electronics*, McGraw-Hill, Inc, New York (1995).
- R. W. Blevins, R. C. Daly, and S. R. Turner, in *Encyclopedia of Polymer Science and Engineering*, Ed. J. I. Krocehwitz, Wiley, New York (1985).
- M. J. Bowden, in *Materials for Microlithography: Radiation-Sensitive Polymers*, Ed. L. F. Thompson, C. G. Willson, and J. M. J. Frechet, American Chemical Society Symposium Series No. 266, Washington, D.C. (1984).
- S. J. Moss and A. Ledwith, *The Chemistry of the Semiconductor Industry*, Blackie & Son Limited, Glasgow (1987).
- E. Reichmanis, F. M. Houlihan, O. Nalamasu, and T. X. Neenan, in *Polymers for Microelectronics*, Ed. L. F. Thompson, C. G. Willson, and S. Tagawa, American Chemical Society Symposium Series, No. 537, Washington, D.C. (1994).
- P. van Zant, *Microchip Fabrication*, 2nd ed., McGraw-Hill Publishing Company, New York (1990).
- C. Grant Willson, in *Introduction to Microlithography*, 2nd ed., Ed. L. F. Thompson, C. G. Willson, M. J. Bowden, American Chemical Society, Washington, D.C. (1983).

Integrated Circuit Well and Gate Creation

Note: This module is based upon the Connexions module entitled *Integrated Circuit Well and Gate Creation* by Bill Wilson.

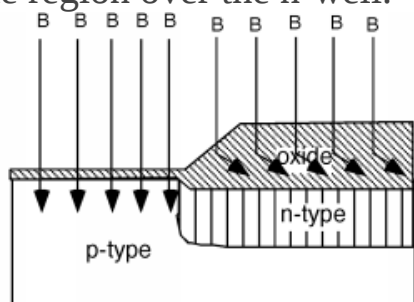
We then remove the remaining resist, and perform an activation/anneal/diffusion step, also sometimes called the "drive-in". The purpose of this step is two fold. We want to make the n-tank deep enough so that we can use it for our p-channel MOS, and we want to build up an implant barrier so that we can implant into the p-substrate region only. We introduce oxygen into the reactor during the activation, so that we grow a thicker oxide over the region where we implanted the phosphorus. The nitride layer over the p-substrate on the LHS protects that area from any oxide growth. We then end up with the structure shown in [\[link\]](#).



After the anneal/drive-in.

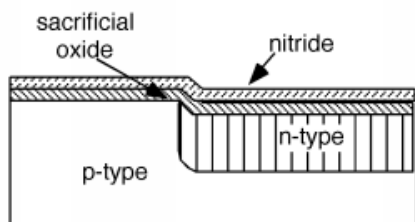
Now we strip the remaining nitride. Since the only way we can convert from p to n is to add a donor concentration which is greater than the background acceptor concentration, we had to keep the doping in the substrate fairly light in order to be able to make the n-tank. The lightly doped p-substrate would have too low a threshold voltage for good n-MOS transistor operation, so we will do a V_T adjust implant through the thin oxide on the LHS, with the thick oxide on the RHS blocking the boron from getting into the n-tank. This is shown in [\[link\]](#), where boron is implanted

into the p-type substrate on the LHS, but is blocked by the thick oxide in the region over the n-well.



V_T adjust implant.

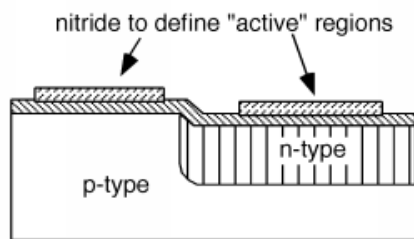
Next, we strip off all the oxide, grow a new thin layer of oxide, and then a layer of nitride [\[link\]](#). The oxide layer is grown only because it is bad to grow Si_3N_4 directly on top of silicon, as the different coefficients of thermal expansion between the two materials causes damage to the silicon crystal structure. Also, it turns out to be nearly impossible to remove nitride if it is deposited directly on to silicon.



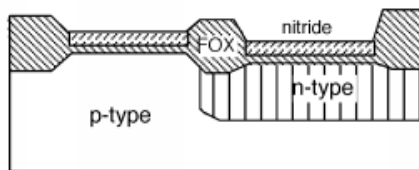
Strip of the oxide and
grow a new nitride layer.

The nitride is patterned (covered with photoresist, exposed, developed, etched, and removal of photoresist) to make two areas which are called "active" [\[link\]](#). The wafer is then subjected to a high-pressure oxidation step which grows a thick oxide wherever the nitride was removed. The nitride is a good barrier for oxygen, so essentially no oxide grows underneath it. The

thick oxide is used to isolate individual transistors, and also to make for an insulating layer over which conducting patterns can be run. The thick oxide is called field oxide (or FOX for short) [\[link\]](#).

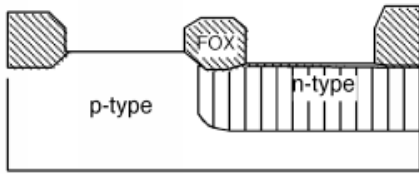


Nitride remaining after etching.



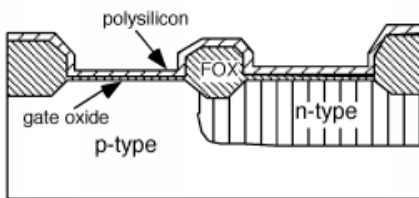
After growth of the field oxide (FOX).

Then, the nitride, and some of the oxide are etched off. The oxide is etched enough so that all of the oxide under the nitride regions is removed, which will take a little off the field oxide as well. This is because we now want to grow the gate oxide, which must be very clean and pure [\[link\]](#). The oxide under the nitride is sometimes called a *sacrificial oxide*, because it is sacrificed in the name of ultra performance.



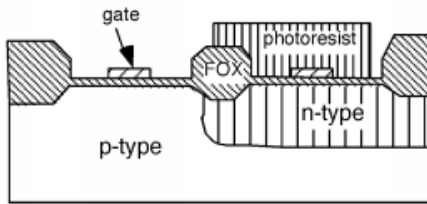
Ready to grow gate
oxide.

Then the gate oxide is grown, and immediately thereafter, the whole wafer is covered with polysilicon [\[link\]](#).



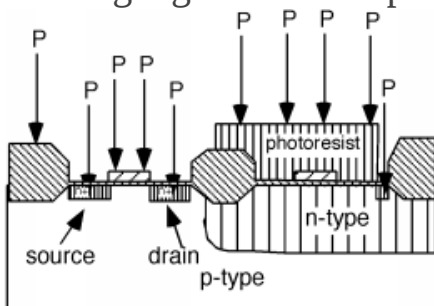
Polysilicon deposition
over the gate oxide.

The polysilicon is then patterned to form the two regions which will be our gates. The wafer is covered once again with photoresist. The resist is removed over the region that will be the n-channel device, but is left covering the p-channel device. A little area near the edge of the n-tank is also uncovered [\[link\]](#). This will allow us to add some additional phosphorus into the n-well, so that we can make a contact there, so that the n-well can be connected to V_{dd} .



Preparing for NMOS
channel/drain implant.

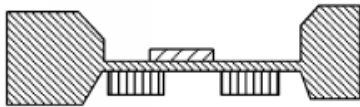
Back into the implanter we go, this time exposing the wafer to phosphorus. The poly gate, the FOX and the photoresist all block phosphorus from getting into the wafer, so we make two n-type regions in the p-type substrate, and we have made our n-channel MOS source/drain regions. We also add phosphorous to the V_{dd} contact region in the n-well so as the make sure we get good contact performance there [\[link\]](#).



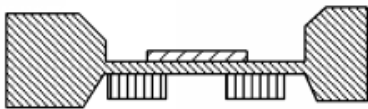
Phosphorus S/D implant.

The formation of the source and drain were performed with a *self-aligning technology*. This means that we used the gate structure itself to define where the two inside edges of the source and drain would be for the MOSFET. If we had made the source/drain regions before we defined the gate, and then tried to line the gate up right over the space between them, we might have gotten something that looks like what is shown in [\[link\]](#). What's going to be the problem with this transistor? Obviously, if the gate

does not extend all the way to both the source and the drain, then the channel will not either, and the transistor will never turn on! We could try making the gate wider, to ensure that it will overlap both active areas, even if it is slightly misaligned, but then you get a lot of extraneous fringing capacitance which will significantly slow down the speed of operation of the transistor [\[link\]](#). This is bad! The development of the self-aligned gate technique was one of the big breakthroughs which has propelled us into the VLSI and ULSI era.

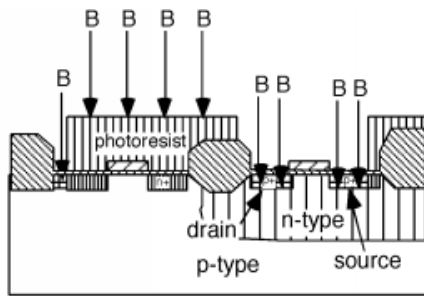


A representation of a misaligned gate.



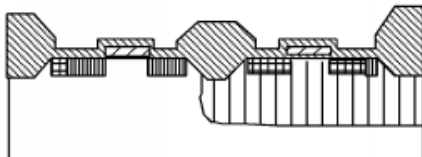
A representation of a wide gate.

We pull the wafer out of the implanter, and strip off the photoresist. This is sometimes difficult, because the act of ion implantation can "bake" the photoresist into a very tough film. Sometimes an rf discharge in an O₂ atmosphere is used to "ash" the photoresist, and literally burn it off the wafer! We now apply some more PR, and this time pattern to have the moat area, and a substrate contact exposed, for a boron p⁺ implant. This is shown in [\[link\]](#).

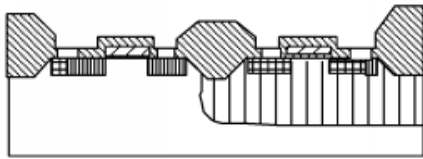


Boron p-channel S/D
implant.

We are almost done. The next thing we do is remove all the photoresist, and grow one more layer of oxide, which covers everything, as shown in [\[link\]](#). We put photoresist over the whole wafer again, and pattern for contact holes to go through the oxide. We will put contacts for the two drains, and for each of the sources, make sure that the holes are big enough to also allow us to connect the source contact to either the p-substrate or the n-moat as is appropriate [\[link\]](#).



Final oxide growth.



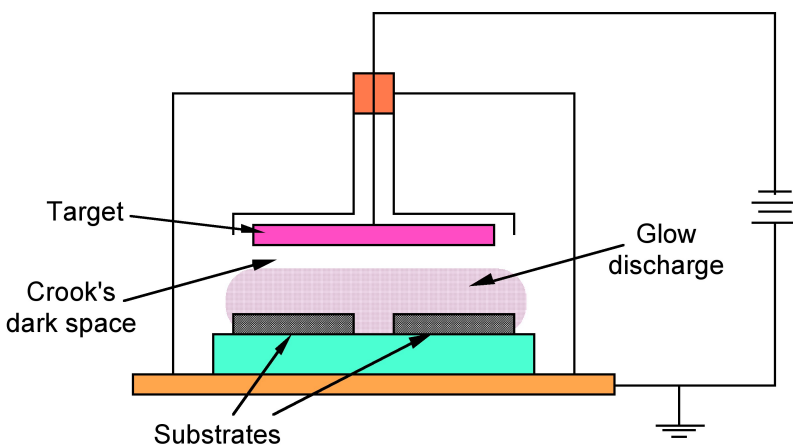
After the contact holes
are etched.

Applying Metallization by Sputtering

Note: This module is adapted from the Connexions module entitled *Applying Metal/Sputtering* by Bill Wilson.

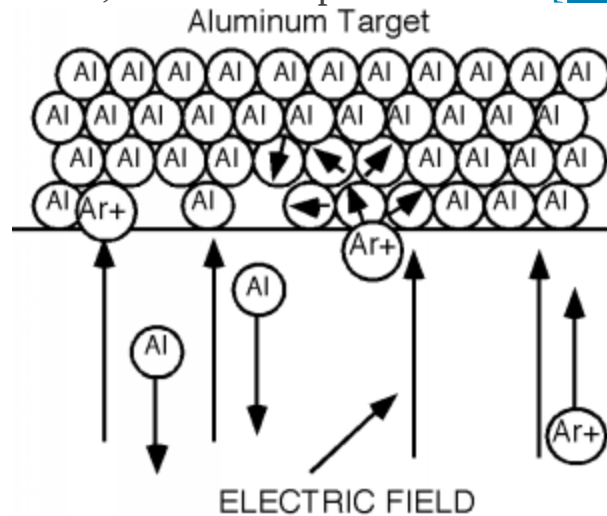
We now put the wafer in a sputter deposition system. In the sputter system, we coat the entire surface of the wafer with a conductor. An aluminum-silicon alloy is usually used, although other metals are employed as well.

A sputtering system is shown schematically in [\[link\]](#). A sputtering system is a vacuum chamber, which after it is pumped out, is re-filled with a low-pressure argon gas. A high voltage ionizes the gas, and creates what is known as the Crookes dark space near the cathode, which in our case, consists of a metal target made out of the metal we want to deposit. Almost all of the potential of the high-voltage supply appears across the dark space. The glow discharge consists of argon ions and electrons which have been stripped off of them. Since there are about equal number of ions and electrons, the net charge density is about zero, and hence by Gauss' law, so is the field.

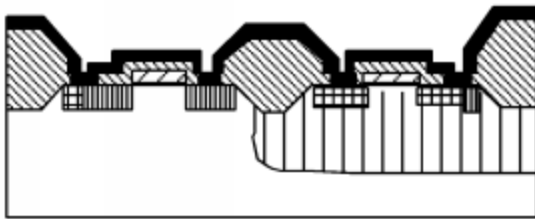


A schematic representation of a sputtering apparatus.

The electric field accelerates the argon atoms which slam into the aluminum target. There is an exchange of momentum, and an aluminum atom is ejected from the target ([\[link\]](#)) and heads to the silicon wafer, where it sticks, and builds up a metal film [\[link\]](#).

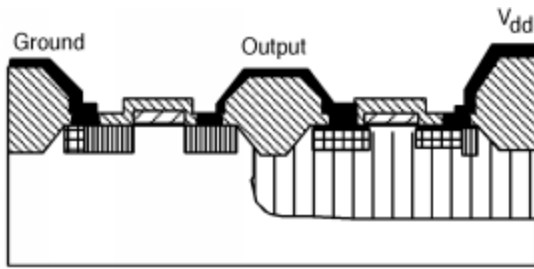


The sputtering mechanism.



Wafer coated with metal.

If you look at [\[link\]](#), you will note that we have seemingly done something pretty stupid. We have wired all of the elements of our CMOS inverter together; but all is not lost. We can do one more photolithographic step, and pattern and etch the aluminum, so we only have it where we need it. This is shown in [\[link\]](#).



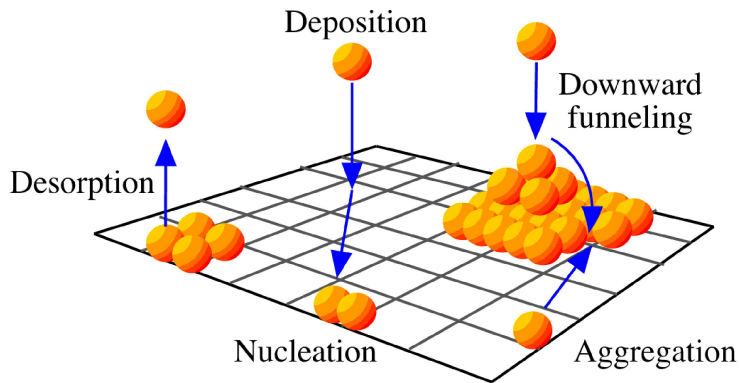
After interconnect patterning.

Molecular Beam Epitaxy

Note: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Sarah Westcott.

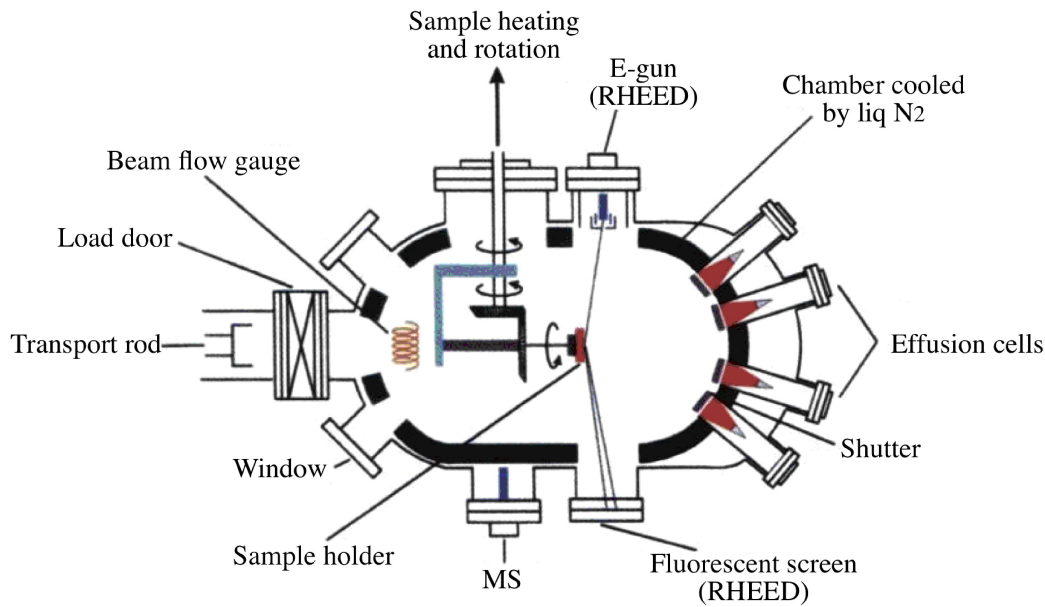
Introduction

In the process of epitaxy, a thin layer of material is grown on a substrate. With respect to crystal growth it applies to the process of growing thin crystalline layers on a crystal substrate. In epitaxial growth, there is a precise crystal orientation of the film in relation to the substrate. For electronic devices, the substrate is a single crystal (usually Si or GaAs) and therefore so is the epitaxial layer (epilayer). In the most basic form of molecular beam epitaxy (MBE), the substrate is placed in ultra high vacuum (UHV) and the source materials for the film are evaporated from elemental sources. The evaporated molecules or atoms flow as a beam, striking the substrate, where they are adsorbed on the surface. Once on the surface, the atoms move by surface diffusion until they reach a thermodynamically favorable location to bond to the substrate. Molecules will dissociate to atomic form during diffusion or at a favorable site. [\[link\]](#) illustrates the processes that can occur on the surface. Because the atoms require time for surface diffusion, the quality of the film will be better with slower growth. Typically growth rates of about 1 monolayer per second provide sufficiently high quality.



Schematic illustration of processes on growing surface during MBE. Adsorption of atoms on the surface, surface diffusion of atoms, formation of crystalline lattice, desorption of particles from the surface.

A typical MBE chamber is shown in [\[link\]](#). The substrate is chemically washed and then put into a loading chamber where it is further cleaned using argon ion bombardment followed by annealing. This removes the top layers of the substrate, which is usually an undesired oxide which grew in air and contains impurities. The annealing heals any damage caused by the bombardment. The substrate then enters the growth chamber via the sample exchange load lock. It is secured on a molybdenum holder either mechanically or with melted indium or gallium which hold the substrate by surface tension.



The MBE growth chamber.

Each effusion cell (see [\[link\]](#)) is a source of one element in the film. The effusion cell, also called a Knudsen cell, contains the elemental form in very high purity (greater than 99.99999% for Ga and As). The cell is heated to encourage evaporation. For GaAs growth, the temperature is typically controlled for a vapor pressure of 10^{-2} to 10^{-3} Torr inside the effusion cell, which results in a transport of about 10^{15} molecules/cm² to the substrate when the shutter for that cell is opened. The shape and size of the opening in the cell is optimized for an even distribution of particles on the substrate. Due to the relatively low concentration of molecules, they typically do not interact with other molecules in the beam during the 5 - 30 cm journey to the substrate. The substrate is usually rotated, at a few rpm, to further even the distribution.

Because MBE takes place in UHV and has relatively low pressure of residual gas at the surface, analysis techniques such as reflection high energy diffraction and ellipsometry can be used during growth, both to study and control the growth process. The UHV environment also allows pre or post growth analysis techniques such as Auger spectroscopy.

Elemental and molecular sources

The effusion cell is used for the majority of MBE growth. All materials used in the cell are carefully chosen to be noninteracting with the element being evaporated. For example, the crucible is pyrolytic boron nitride. However, it has disadvantages, such as:

- The evaporated species may be molecular, rather than monomeric, which will require further dissociation at the surface.
- When the shutter is opened, the heat loss from the cell results in a transient in the beam flux which last for several minutes and cause variations of up to 50%.
- The growth chamber must be opened up to replace the solid sources.

Cracker cells are used to improve the ratio of monomeric to molecular (or at least dimeric to tetrameric) particles from the source. The cracker cell, placed so that the beam passes through it after the effusion cell, is maintained at a high temperature (and sometimes high pressure) to encourage dissociation. The dissociation process generally requires a catalyst and the best catalysts for a given species have been studied.

Some elements, such as silicon, have low enough vapor pressure that more direct heating techniques such as electron bombardment or laser radiation heating are used. The electron beam is bent using electromagnetic focusing to prevent any impurities in the electron source from contaminating the silicon to be used in MBE. Because the heat is concentrated on the surface to be evaporated, interactions with and contamination from the crucible walls is reduced. In addition, this design does not require a shutter, so there is no problem with transients. Modulation of the beam can produce very sharp interfaces on the substrate. In laser radiation heating, the electron beam is replaced by a laser beam. The advantages of localized heating and rapid modulation are also maintained without having to worry about contamination from the electron source or stray electrons.

Some of the II-VI (12-16) compounds have such high vapor pressure that a Knudsen cell cannot be used. For example, the mercury source must be kept cooler than the substrate to keep the vapor pressure low enough to be

feasible. The Hg source must also be sealed off from the growth chamber to allow the chamber to be pumped down.

Two other methods of obtaining the elements for use in epitaxy are gas-source epitaxy and chemical beam epitaxy (CBE). Both of these methods use gas sources, but they are distinguished by the use of elemental beams in gas source epitaxy, while organometallic beams are used in CBE. For the example of III-V (13-15) semiconductors, in gas epitaxy, the group III material may come from an effusion cell while the group V material is the hydride, such as AsH_3 or PH_3 , which is cracked before entering the growth chamber. In CBE, the group V material is an organometallic, such as triethylgallium $[\text{Ga}(\text{C}_2\text{H}_5)_3]$ or trimethylaluminum $[\text{Al}(\text{CH}_3)_3]$, which adsorbs on the surface, where it dissociates.

The gas sources have several advantages. Gas lines can be run into the chamber, which allows the supply to be replenished without opening the chamber. When making alloys, such as $\text{Al}_x\text{Ga}_{1-x}\text{As}$, the gases can be premixed for the correct stoichiometry or even have their composition gradually changed for making graded structures. For abrupt structures, it is necessary to be able to switch the gas lines with speeds of 1 second or less. However, the gas lines increase the complexity of the process and can be hard on the pumping system.

Substrate choice and preparation

Materials can be grown on substrates of different structure, orientation, and chemistry. In deciding which materials can be grown on a particular substrate, a primary consideration was expected to be lattice mismatch, i.e., differences in spacing between atoms. However, while lattice mismatch can cause strain in the grown layer, considerable accommodation between materials of different sizes can take place during growth. A greater source of strain can be differences in thermal expansion characteristics because the layer is grown at high temperature. On cooling to room temperature, dislocation defects can be formed at the interface or in severe cases, the device may break. Chemical considerations, such as whether the layer's elements will dissolve in the substrate or form compounds with the substrate, must also be considered.

Different orientations of the substrate can also affect growth. Close-packed planes have the lowest surface energy, which allows atoms to desorb from the surface, resulting in slower growth rates. Growth is favored where bonds can be made in several directions at the same time. Therefore, the substrate is often cut off-axis by a $2 - 4^\circ$ to provide a rougher growth surface. For compound semiconductors, some orientations result in the number of loose bonds changing between layers. This results in changes of surface energy with each layer, which slows growth down.

The greatest cause of defects in the epitaxial layer is defects on the substrate's surface. In general, any dislocations on the substrate are replicated or even multiplied in the epitaxial growth, which is what makes the cleaning of the substrate so important.

Materials grown

MBE is commercially used primarily for GaAs devices. This is partly because the high speed microwave devices made from GaAs required the superior electrical quality of epitaxial layers. Taking place at lower temperature and under better controlled conditions, MBE generally results in layers of better quality than melt-grown.

From solid Ga and As sources, elemental Ga and tetrameric As_4 are evaporated. For a GaAs substrate, the Ga flux has a sticking coefficient very close to 1 (almost certain to adsorb). The As is much less likely to adsorb, so an excess is usually supplied. Cracker cells are often used on the As_4 in order to obtain As_2 instead, which results in faster growth and more efficient utilization of the source beam.

For nominally undoped GaAs grown by MBE, the residual impurities are usually carbon, from substrate contamination and residual gas after the growth chamber is pumped down, and sulphur, from the As source. The most common surface defects are "oval" defects, which seem to form when Ga manages to form metallic droplets during the growth process, which can particularly occur if the substrate was not cleaned properly. These defects can also be reduced by carefully controlling the Ga flux.

During MBE growth, dopants can be introduced by having a separate effusion cell or gas source for each dopant. To achieve a desired dopant concentration in the film, not only must the rate of dopants striking the substrate be controlled, but the characteristics of how the dopant behaves on the surface must be known. Low-vapor pressure dopants tend to desorb from the surface and their behavior is very temperature dependent and so are avoided when possible. Slow diffusing dopants adsorb to surface sites and are eventually covered as more GaAs is grown. Their incorporation depends linearly on the partial pressure of the dopant present in the growth chamber. This is the behavior exhibited by most n-type dopants in GaAs and most dopants of both types in Si. If the dopant, like most p-type GaAs dopants, is able to diffuse through the surface of the substrate into the crystal below, then there will be higher incorporation efficiency, which will depend on the square root of the dopant partial pressure for reasonable concentrations. Due to increasing lattice strain, all dopants will saturate at very high concentrations. They may also tend to form clusters. Dopant behavior depends on many factors and is actively studied.

The growth of GaAs epitaxial layers on silicon substrates has also been investigated. Silicon substrates are grown in larger wafers, have better thermal conductivity allowing more devices/chip to be grown on them, and are cheaper. However, because Si is nonpolar and GaAs is polar, the GaAs tends to form islands on the surface with different phase (what should be a Ga site based on a neighboring domain's pattern will actually be an As site). There is also a fairly large lattice mismatch, leading to many dislocations. However, FETs, LEDs, and lasers have all been made in laboratories.

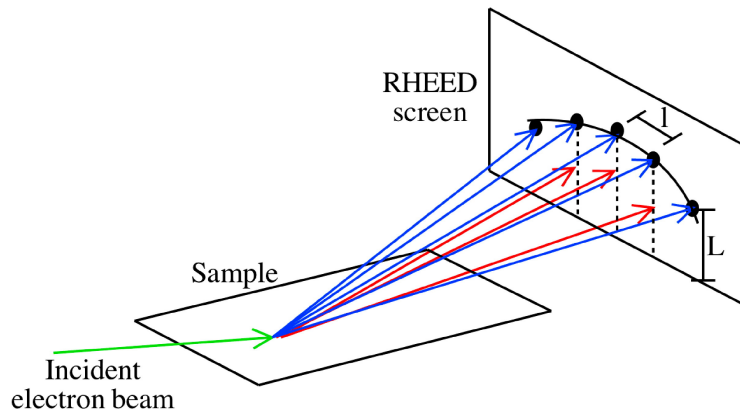
Many devices require abrupt junctions between layers of different materials. One group, studying how to make high quality, abrupt GaAs and AlAs layers, found that rapid movement of the Ga or Al on the surface was required. This migration was enhanced at high temperatures, but unfortunately, diffusion into the substrate also increased. However, they also discovered that migration of Ga or Al increased if the As supply was turned off. By alternating the Ga and As supplies, the Ga was able to reach the substrate and migrate to provide more even monolayer coverage before the As atoms arrived to react.

Besides GaAs, most other III-V semiconductors have also been grown using MBE. Structures involving very thin layers (only a few atomic layers thick), often called superlattices or strained superlattices if there is a large lattice mismatch, are routinely grown. Because different materials have different energy levels for electrons and holes, it is possible to trap carriers in one of these thin layers, forming a quantum well. This type of confinement structure is particularly popular for LEDs or lasers, including blue light lasers. The strained superlattice structure actually shifts and splits the energy levels of the materials in some cases making devices possible for such applications as infrared light detection, which requires very small band gaps.

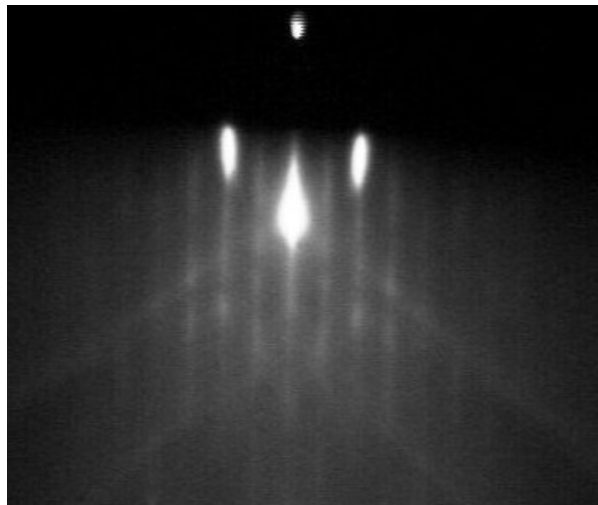
Thin films of many other materials have also been grown using MBE methods. Silicon technology has cheaper methods of growth and so Si layers are not very popular. However, possible devices made of Si-Ge alloys have been grown. The II-VI compounds, have also been grown. Magnetic materials, such as Co-Pt and Fe-Pt alloys, have been grown in the hopes of providing better magnetic storage.

Analysis techniques

The most popular in-situ analysis technique for MBE-grown layers is reflection high energy diffraction (RHEED), see [\[link\]](#). Electrons of energy 5 - 40 keV are directed towards the sample. They reflect from the surface at a very small angle (less than 3°) and are directed onto a screen. These electrons interact with only the top few atomic layers and thus provide information about the surface. [\[link\]](#) shows a typical pattern on the screen for electrons reflected from a smooth surface, in which constructive interference between some of the electrons reflected from the lattice structure results in lines. If the surface is rough, spots will appear on the screen, also. By looking at the total intensity of the reflected electron pattern, an idea of the number of monolayers deposited and how epilayers grow can be obtained. The island-type growth shown in this figure is an area of intense interest. These oscillations in intensity are gradually damped as more layers are grown, because the overall roughness of the surface increases.



Schematic illustrating the formation of a RHEED pattern.



RHEED diffraction pattern of a GaAs surface. Adapted from images by the MBE Laboratory in the Institute of Physics of the ASCR
 (<http://www.fzu.cz/departments/surfaces/mbe/index.html>)

Phase locked epitaxy takes advantage of the patterns of the oscillations to grow very abrupt layers. By sending the oscillation information to a computer, it can decide when to open or close the shutters of the effusion cell based on the location in the oscillation cycle. This technique self-adjusts for fluctuations in beam flux when the shutters are opened and can grow very abrupt layers.

Another analysis technique that can be used to study surface smoothness during growth is ellipsometry. Polarized laser light is reflected from the surface at a small angle. The polarization of the light changes, depending on the roughness of the surface.

Improved growth characteristics also require that the actual flux from the sources is measured. This is typically done with an ion gauge flux monitor, which is either used to measure residual beam that misses the substrate or is moved into the beam path for calibration when a new source is used. Because of the importance of clean substrate surfaces for low-defect growth, Auger spectroscopy is used following cleaning by sputtering. Auger spectroscopy takes place by ionizing an inner shell electron from an atom. When an outer shell electron then deexcites to the inner shell, the energy released can prompt the emission of another outer shell electron. The energy at which this occurs is characteristic of the atom involved and the signal can be used to detect impurities as small as 0.1%.

Bibliography

- K. J. Bachmann, *The Materials Science of Microelectronics*, VCH (1995).
- S. K. Ghandhi, *VLSI Fabrication Principles: Silicon and Gallium Arsenide*, 2nd Edition, Wiley-Interscience, NY (1994).
- M. A. Herman and H. Sitter, *Molecular Beam Epitaxy: Fundamentals and Current Status*, Springer-Verlag (1989).
- Y. Horikoshi, M. Kawashima, and H. Yamaguchi, *Jpn. J. Appl. Phys.*, 1986, **25**, L868.
- J. H. McFee, B. I. Miller, and K. J. Bachmann, *J. Electrochem. Soc.*, 1977, **124**, 259.

- T. Sakamoto, H. Funabashi, K. Ohta, T. Nakagawa, N. J. Kawai, and T. Kojima, *Jpn. J. Appl. Phys.*, 1984, **23**, L657.
- W. T. Tsang, *J. Crystal Growth*, 1987, **81**, 261.

Atomic Layer Deposition

This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Julie A. Francis.

Introduction

The growth of thin films has had dramatic impact on technological progress. Because of the various properties and functions of these films, their applications are limitless especially in microelectronics. These layers can act as superconductors, semiconductors, conductors, insulators, dielectric, or ferroelectrics. In semiconductor devices, these layers can act as active layers and dielectric, conducting, or ion barrier layers. Depending on the type of film material and its applications, various deposition techniques may be employed. For gas-phase deposition, these include vacuum evaporation, reactive sputtering, chemical vapor deposition (CVD), especially metal organic CVD (MOCVD), and molecular beam epitaxy (MBE). Atomic layer deposition (ALD), originally called atomic layer epitaxy (ALE), was first reported by Suntola et al. in 1980 for the growth of zinc sulfide thin films to fabricate electroluminescent flat panel displays.

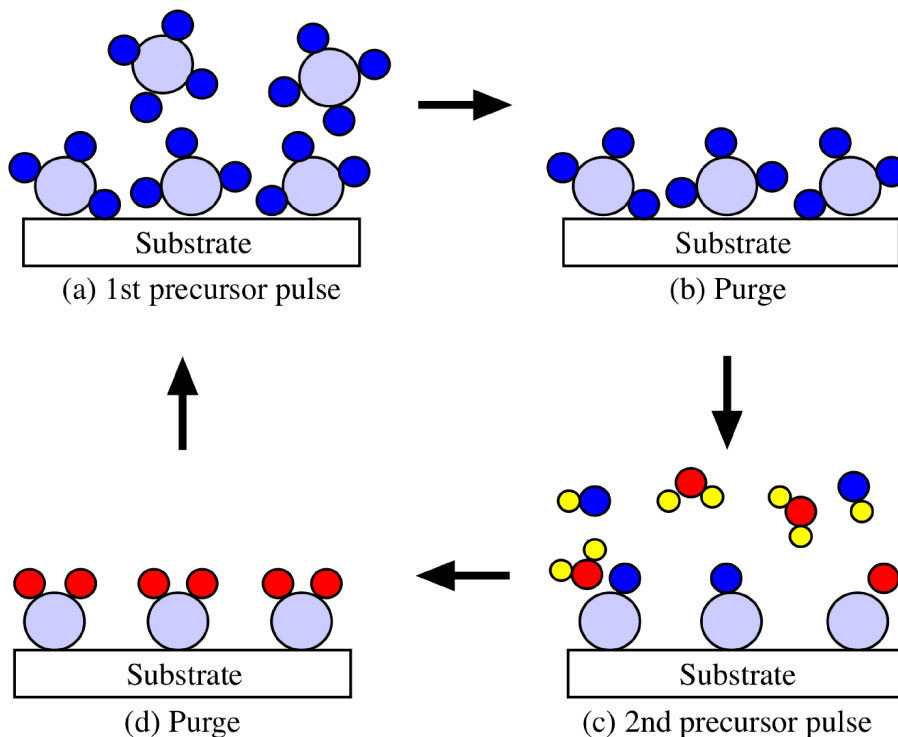
ALD refers to the method whereby film growth occurs by exposing the substrate to its starting materials alternately. It should be noted that ALE is actually a sub-set of ALD, in which the grown film is epitaxial to the substrate; however, the terms are often used interchangeably. Although both ALD and CVD use chemical (molecular) precursors, the difference between the techniques is that the former uses self limiting chemical reactions to control in a very accurate way the thickness and composition of the film deposited. In this regard ALD can be considered as taking the best of CVD (the use of molecular precursors and atmospheric or low pressure) and MBE (atom-by-atom growth and a high control over film thickness) and combining them in single method. A selection of materials deposited by ALD is given in [\[link\]](#).

Compound class	Examples
II–VI compounds	ZnS, ZnSe, ZnTe, $\text{ZnS}_{1-x}\text{Se}_x$, CaS, SrS, BaS, $\text{SrS}_{1-x}\text{Se}_x$, CdS, CdTe, MnTe, HgTe, $\text{Hg}_{1-x}\text{Cd}_x\text{Te}$, $\text{Cd}_{1-x}\text{Mn}_x\text{Te}$
II–VI based thin-film electroluminescent (TFEL) phosphors	ZnS:M (M = Mn, Tb, Tm), CaS:M (M = Eu, Ce, Tb, Pb), SrS:M (M = Ce, Tb, Pb, Mn, Cu)
III–V compounds	GaAs, AlAs, AlP, InP, GaP, InAs, $\text{Al}_x\text{Ga}_{1-x}\text{As}$, $\text{Ga}_x\text{In}_{1-x}\text{As}$, $\text{Ga}_x\text{In}_{1-x}\text{P}$
Semiconductors/dielectric nitrides	AlN, GaN, InN, SiN_x
Metallic nitrides	TiN, TaN, Ta_3N_5 , NbN, MoN
Dielectric oxides	Al_2O_3 , TiO_2 , ZrO_2 , HfO_2 , Ta_2O_5 , Nb_2O_5 , Y_2O_3 , MgO, CeO_2 , SiO_2 , La_2O_3 , SrTiO_3 , BaTiO_3
Transparent conductor oxides	In_2O_3 , $\text{In}_2\text{O}_3\text{:Sn}$, $\text{In}_2\text{O}_3\text{:F}$, $\text{In}_2\text{O}_3\text{:Zr}$, SnO_2 , $\text{SnO}_2\text{:Sb}$, ZnO,
Semiconductor oxides	ZnO:Al , Ga_2O_3 , NiO, CoO_x
Superconductor oxides	$\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$
Fluorides	CaF_2 , SrF_2 , ZnF_2

Examples of thin film materials deposited by ALD including films deposited in epitaxial, polycrystalline or amorphous form. Adapted from M. Ritala and M. Leskel, *Nanotechnology*, 1999, **10**, 19.

How ALD works

The premise behind the ALD process is a simple one. The substrate (amorphous or crystalline) is exposed to the first gaseous precursor molecule (elemental vapor or volatile compound of the element) in excess and the temperature and gas flow is adjusted so that only one monolayer of the reactant is chemisorbed onto the surface ([link](#)a). The excess of the reactant, which is in the gas phase or physisorbed on the surface, is then purged out of the chamber with an inert gas pulse before exposing the substrate to the other reactant ([link](#)b). The second reactant then chemisorbs and undergoes an exchange reaction with the first reactant on the substrate surface ([link](#)c). This results in the formation of a solid molecular film and a gaseous side product that may then be removed with an inert gas pulse ([link](#)d).



Schematic representation of an ALD process.

The deposition may be defined as self-limiting since one, and only one, monolayer of the reactant species remains on the surface after each exposure. In this case, one complete cycle results in the deposition of one monolayer of the compound on the substrate. Repeating this cycle leads to a controlled layer-by-layer growth. Thus the film thickness is controlled by the number of precursor cycles rather than the deposition time, as is the case for a CVD processes. This self-limiting behavior is the fundamental aspect of ALD and understanding the underlying mechanism is necessary for the future exploitation of ALD.

One basic condition for a successful ALD process is that the binding energy of a monolayer chemisorbed on a surface is higher than the binding energy of subsequent layers on top of the formed layer; the temperature of the reaction controls this. The temperature must be kept low enough to keep the monolayer on the surface until the reaction with the second reactant occurs, but high enough to re-evaporate or break the chemisorption bond. The control of a monolayer can further be influenced with the input of extra energy such as UV irradiation or laser beams. The greater the difference between the bond energy of a monolayer and the bond energies of the subsequent layers, the better the self-controlling characteristics of the process.

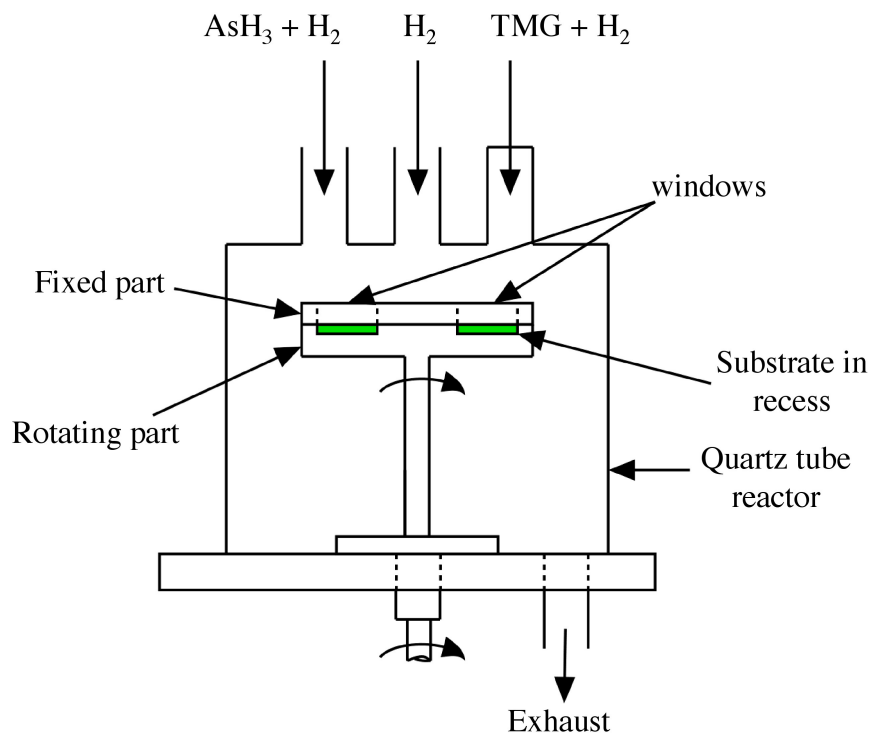
Basically, the ALD technique depends on the difference between chemisorption and physisorption. Physisorption involves the weak van der Waal's forces, whereas chemisorption involves the formation of relatively strong chemical bonds and requires some activation energy, therefore it may be slow and not always reversible. Above certain temperatures chemisorption dominates and it is at this temperature ALD operates best. Also, chemisorption is the reason that the process is self-controlling and insensitive to pressure and substrate changes because only one atomic or molecular layer can adsorb at the same time.

Equipment for the ALD process

Equipment used in the ALD process may be classified in terms of their working pressure (vacuum, low pressure, atmospheric pressure), method of

pulsing the precursors (moving substrate or valve sources) or according to the types of sources. Several system types are discussed.

In a typical moving substrate ALD growth system ([\[link\]](#)) the substrate, located in the recess part of the susceptor, is continuously rotated and cuts through streams of the gaseous precursors, in this case, trimethylgallium [TMG, $\text{Ga}(\text{CH}_3)_3$] and arsine (AsH_3). These gaseous precursors are introduced through separate lines and the gases come in contact with the substrate only when it revolves under the inlet tube. This cycle is repeated until the required thickness of GaAs is achieved. The exposure time to each of the gas streams is about 0.3 s.



A typical moving substrate ALD growth system used to grow GaAs films. Adapted from M. A. Tischler and S. M. Bedair, *Appl. Phys. Lett.*, 1986, **48**, 1681.

ALD may be carried out in a vacuum system using an ultra-high vacuum with a movable substrate holder and gaseous valving. In this manner it may be also equipped with an in-situ LEED system for the direct observation of surface atom configurations and other systems such as XPS, UPS, and AES for surface analysis.

A lateral flow system may also be employed for successful ALE deposition. This uses an inert gas flow for several functions; it transports the reactants, it prevents pump oil from entering the reaction zone, it valves the sources and it purges the deposition site between pulses. Inert gas valving has many advantages as it can be used at ultra high temperatures where mechanical valves may fail and it does not corrode as mechanical valves would in the presence of halides. This method is based on the fact that as the inert gas is flowing through the feeding tube from the source to the reaction chamber, it blocks the flow of the sources. Although in this system the front end of the substrate receives a higher flux density than the down-stream end, a uniform growth rate occurs as long as the saturation layer of the monofunctional group predominates. This lateral flow system effectively utilizes the saturation mechanism of a monolayer formation obtained in ALE. Depending on the properties of the precursors used, and on the growth temperature, various growth systems may be used for ALE.

Requirements for ALD growth

Several parameters must be taken into account in order to assure successful ALD growth. These include the physical and chemical properties of the source materials, their pulsing into the reactor, their interaction with the substrate and each other, and the thermodynamics and volatility of the film itself.

Source molecules used in ALD can be either elemental or an inorganic, organic, or organometallic compound. The chemical nature of the precursor is insignificant as long as it possesses the following properties. It must be a gas or must volatilize at a reasonable temperature producing sufficient vapor pressure. The vapor pressure must be high enough to fill the substrate area so that the monolayer chemisorption can occur in a reasonable length of time. Note that prolonged exposure to the substrate can cause the

precursor to condense on the surface hindering the growth. Chemical interaction between the two precursors prior to chemisorption on the surface is also undesired. This may be overcome by purging the surface with an inert gas or hydrogen between the pulses. The inert gas not only separates the reactant pulses but also cleans out the reaction area by removing excess molecules. Also, the source molecules should not decompose on the substrate instead of chemisorbing. The decomposition of the precursor leads to uncontrolled growth of the film and this defeats the purpose of ALD as it no longer is self-controlled, layer-by-layer growth and the quality of the film plummets.

In general, temperature remains the most important parameter in the ALD process. There exists a processing window for ideal growth of monolayers. The temperature behavior of the rate of growth in monolayer units per cycle gives a first indication of the limiting mechanisms of an ALD process. If the temperature falls too low, the reactant may condense or the energy of activation required for the surface reaction may not be attained. If the temperature is too high, then the precursor may decompose or the monolayer may evaporate resulting in poor ALD growth. In the appropriate temperature window, full monolayer saturation occurs meaning that all bonding sites are occupied and a growth rate of one lattice unit per cycle is observed. If the saturation density is below one, several factors may contribute to this. These include an oversized reactant molecule, surface reconstruction, or the bond strength of an adsorbed surface atom is higher when the neighboring sites are unoccupied. Then the lower saturation density may be thermodynamically favored. If the saturation density is above one, then the undecomposed precursor molecules form the monolayer. Generally, ideal growth occurs when the temperature is set where the saturation density is one.

Advantages of ALD

Atomic layer deposition provides an easy way to produce uniform, crystalline, high quality thin films. It has primarily been directed towards epitaxial growth of III-V (13-15) and II-V (12-16) compounds, especially to layered structures such as superlattices and superalloys. This application is due to the greatest advantage of this method, it is controllable to an

accuracy of a single atomic layer because of saturated surface reactions. Not only that, but it produces epitaxial layers that are uniform over large areas, even on non-planar surfaces, at temperatures lower than those used in conventional epitaxial growth.

Another advantage to this method that may be most important for future applications, is the versatility associated with the process. It is possible to grow different thin films by choosing suitable starting materials among the thousands of available chemical compounds. Provided that the thermodynamics are favorable, the adjustment of the reaction conditions is a relatively easy task because the process is insensitive to small changes in temperature and pressure due to its relatively large processing window. There are also no limits in principle to the size and shape of the substrates.

One advantage that is resultant from the self-limiting growth mechanism is that the final thickness of the film is dependent only upon the number of deposition cycles and the lattice constant of the material, and can be reproduced and controlled. The thickness is independent of the partial pressures of the precursors and growth temperature. Under ideal conditions, the uniformity and the reproducibility of the films are excellent. ALE also has the potential to grow very abrupt heterostructures and very thin layers and these properties are in demand for some applications such as superlattices and quantum wells.

Bibliography

- D. C. Bradley, *Chem. Rev.*, 1989, **89**, 1317.
- M. Ritala and M. Leskel, *Nanotechnology*, 1999, **10**, 19.
- M. Pessa, P. Huttunen, and M. A. Herman, *J. Appl. Phys.*, 1983, **54**, 6047.
- T. Suntola and J. Antson, *Method for producing compound thin films*, U.S. Patent 4,058,430 (1977).
- M. A. Tischler and S. M. Bedair, *Appl. Phys. Lett.*, 1986, **48**, 1681.

Chemical Vapor Deposition

Note: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Scott Stokes.

Introduction

Chemical vapor deposition (CVD) is a deposition process where chemical precursors are transported in the vapor phase to decompose on a heated substrate to form a film. The films may be epitaxial, polycrystalline or amorphous depending on the materials and reactor conditions. CVD has become the major method of film deposition for the semiconductor industry due to its high throughput, high purity, and low cost of operation. CVD is also commonly used in optoelectronics applications, optical coatings, and coatings of wear resistant parts.

CVD has many advantages over physical vapor deposition (PVD) processes such as molecular beam evaporation and sputtering. Firstly, the pressures used in CVD allow coating of three dimensional structures with large aspect ratios. Since evaporation processes are very directional, PVD processes are typically line of sight depositions that may not give complete coverage due to shadowing from tall structures. Secondly, high precursor flow rates in CVD give deposition rates several times higher than PVD. Also, the CVD reactor is relatively simple and can be scaled to fit several substrates. Ultra-high vacuum is not needed for CVD and changes or additions of precursors is an easy task. Furthermore, varying evaporation rates make stoichiometry hard to control in physical deposition. While for CVD stoichiometry is more easily controlled by monitoring flow rates of precursors. Other advantages of CVD include growth of high purity films and the ability to fabricate abrupt junctions.

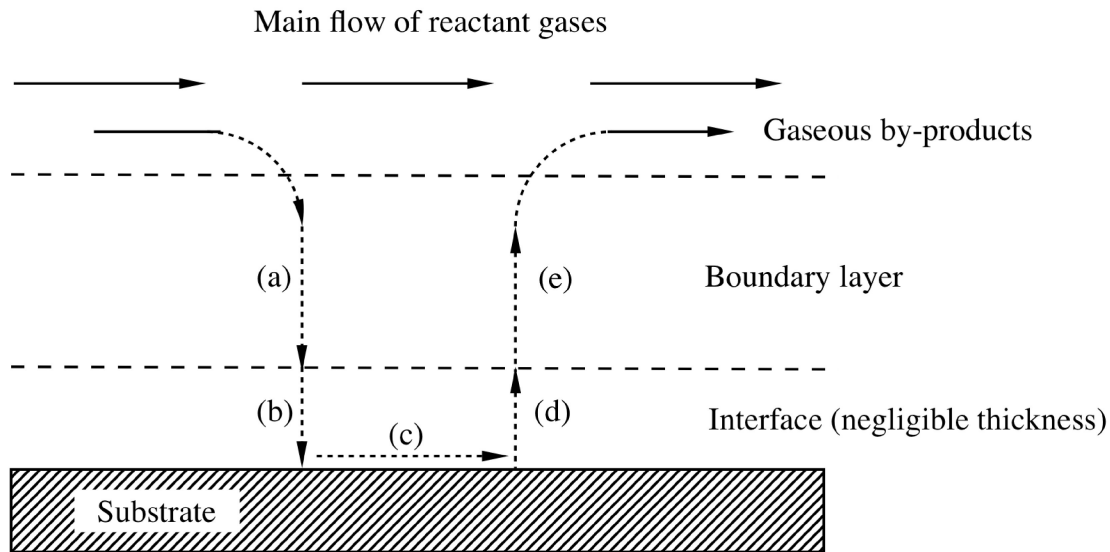
There are, however, some disadvantages of CVD that make PVD more attractive for some applications. High deposition temperatures for some

CVD processes (often greater than 600 °C) are often unsuitable for structures already fabricated on substrates. Although with some materials, use of plasma-enhanced CVD or metal-organic precursors may reduce deposition temperatures. Another disadvantage is that CVD precursors are often hazardous or toxic and the by-products of these precursors may also be toxic. Therefore extra steps have to be taken in the handling of the precursors and in the treatment of the reactor exhaust. Also, many precursors for CVD, especially the metal-organics, are relatively expensive. Finally, the CVD process contains a large number of parameters that must be accurately and reproducibly optimized to produce good films.

Kinetics of CVD

A normal CVD process involves complex flow dynamics since gases are flowing into the reactor, reacting, and then by-products are exhausted out of the reactor. The sequence of events during a CVD reaction are shown in [\[link\]](#) and as follows:

1. Precursor gases input into the chamber by pressurized gas lines.
2. Mass transport of precursors from the main flow region to the substrate through the boundary layer ([\[link\]](#)a);
3. Adsorption of precursors on the substrate (normally heated) ([\[link\]](#)b).
4. Chemical reaction on the surface ([\[link\]](#)c)
5. Atoms diffuse on the surface to growth sites.
6. Desorption of by-products of the reactions ([\[link\]](#)d).
7. Mass transport of by-products to the main flow region ([\[link\]](#)e).



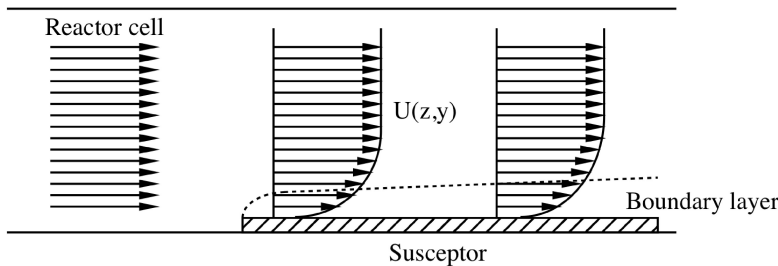
Sequence of events during CVD: (a) diffusion of reactants through boundary layer, (b) adsorption of reactants on substrate, (c) chemical reaction takes place, (d) desorption of adsorbed species, and (e) diffusion out of by-products through boundary layer.

Adapted from H. O. Pierson, *Handbook of Chemical Vapor Deposition*, Noyes Publications, Park Ridge (1992).

The boundary layer

Gas flow in a CVD reactor is generally laminar, although in some cases heating of the chamber walls will create convection currents. The complete problem of gas flow through the system is too complex to be described here; however, assuming we have laminar flow (often a safe assumption) the gas velocity at the chamber walls will be zero. Between the wall (zero velocity) and the bulk gas velocity there is a boundary layer. The boundary layer thickness increases with lowered gas velocity and the distance from the tube inlet ([\[link\]](#)). Reactant gases flowing in the bulk must diffuse through the boundary layer to reach the substrate surface. Often, the

susceptor is tilted to partially compensate for the increasing boundary-layer thickness and concentration profile.

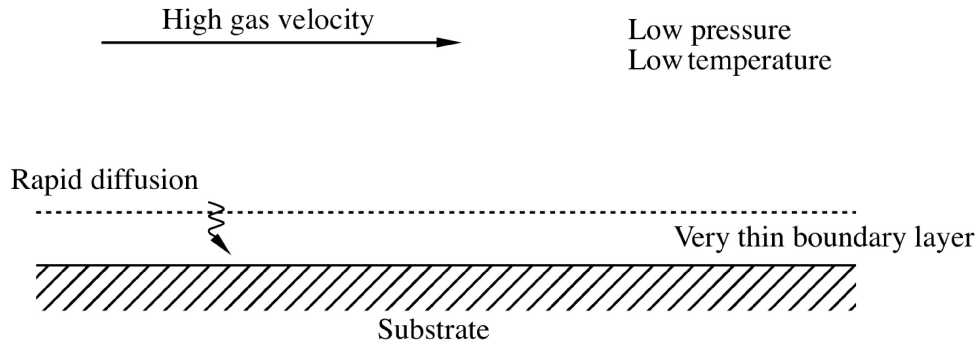


Development of boundary layer in a horizontal reactor. Adapted from G. B. Stringfellow, *Organometallic Vapor-Phase Epitaxy: Theory and Practice*, Academic Press, New York (1994).

Rate limiting steps

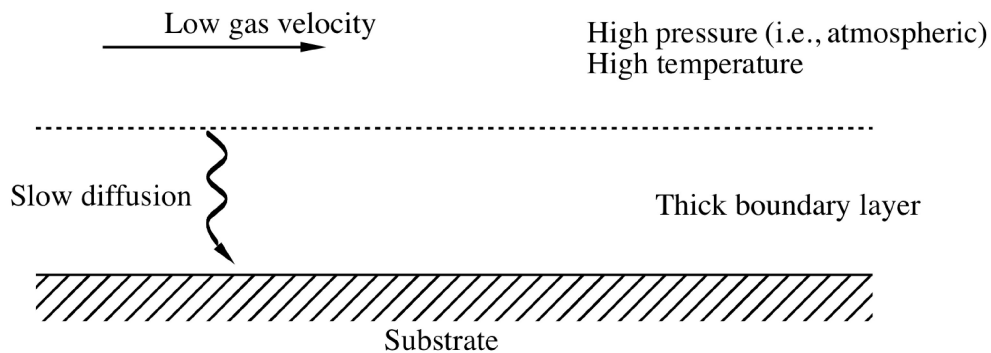
During CVD the growth rate of the film is limited by either surface reaction kinetics, mass transport (diffusion) of precursors to the substrate, or the feed rate of the precursors.

Surface reaction controls the rate when growth occurs at low temperatures (where the reaction occurs slowly) and also dominates at low pressures (where the boundary layer is thin and reactants easily diffuse to the surface), see [\[link\]](#). Since reactants easily diffuse through the boundary layer, the amount of reactant at the surface is independent of reactor pressure. Therefore, it is the reactions and motions of the precursors adsorbed on the surface which will determine the overall growth rate of the film. A sign of surface reaction limited growth would be dependence of the growth rate on substrate orientation, since the orientation would certainly not affect the thermodynamics or mass transport of the system.



Surface reaction limited growth in CVD. Adapted from H. O. Pierson, *Handbook of Chemical Vapor Deposition*, Noyes Publications, Park Ridge (1992).

A deposition limited by mass transport is controlled by the diffusion of reactants through the boundary layer and diffusion of by-products out of the boundary layer. Mass transport limits reactions when the temperature and pressure are high. These conditions increase the thickness of the boundary layer and make it harder for gases to diffuse through ([link](#)). In addition, decomposition of the reactants is typically quicker since the substrate is at a higher temperature. When mass transport limits the growth, either increasing the gas velocity or rotating the substrate during growth will decrease the boundary layer and increase the growth rate.

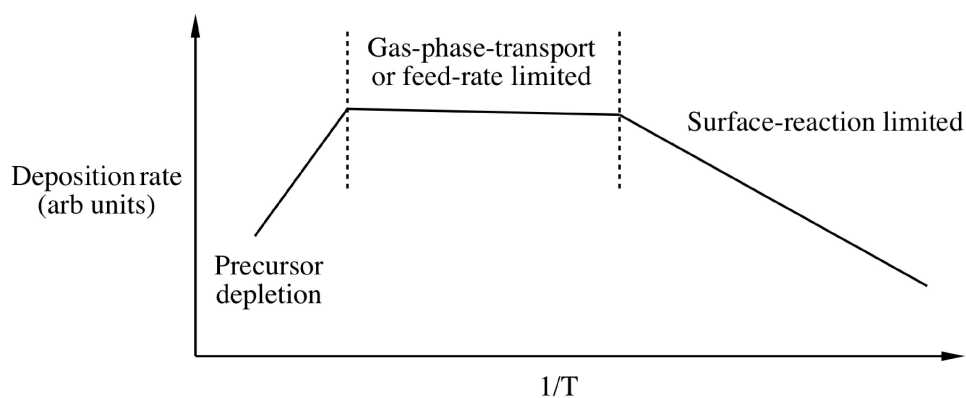


Mass transport limited growth in CVD. Adapted from H.

O. Pierson, *Handbook of Chemical Vapor Deposition*,
Noyes Publications, Park Ridge (1992).

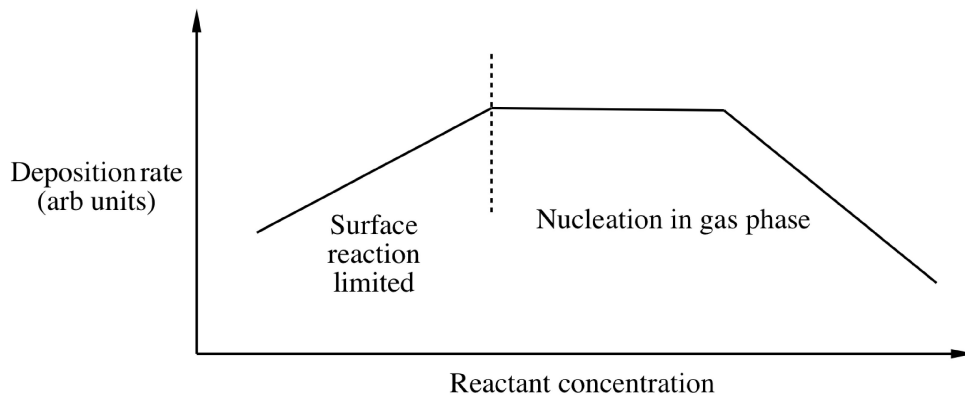
Feed rate limits the deposition when nearly all the reactant is consumed in the chamber. The feed rate is more important for a hot wall reactor since the heated walls will decompose a large amount of the precursor. Cold wall reactors tend to have higher deposition rates since the reactants are not depleted by the walls.

A plot of growth rate versus temperature, known as an Arrhenius plot, can be used to determine the rate limiting step of a reaction ([\[link\]](#)). Mass transport limits reactions at high temperatures such that growth rate increases with partial pressures of reactants, but is constant with temperature. Surface reaction kinetics dominates at low temperatures where the growth rate increases with temperature, but is constant with pressures of reactants. Feed rate limited reactions are independent of temperature, since it is the rate of gas delivery that is limiting the reaction. The Arrhenius plot will show where the transition between the mass transport limited and the surface kinetics limited growth occurs in the temperature regime.



Dependence of CVD deposition rate on temperature.
Adapted from J. G. Eden, in *Thin Film Processes II*, Eds.
J. L. Vossen and W. Kern, Academic Press, New York
(1991).

Increases in reactant concentrations will to a point increase the deposition rate. However, at very high reactant concentrations, gas phase nucleation will occur and the growth rate will drop ([\[link\]](#)). Slow deposition in a CVD reactor can often be attributed to either gas phase nucleation, precursor depletion due to hot walls, thick boundary layer formation, low temperature, or low precursor vapor pressure.



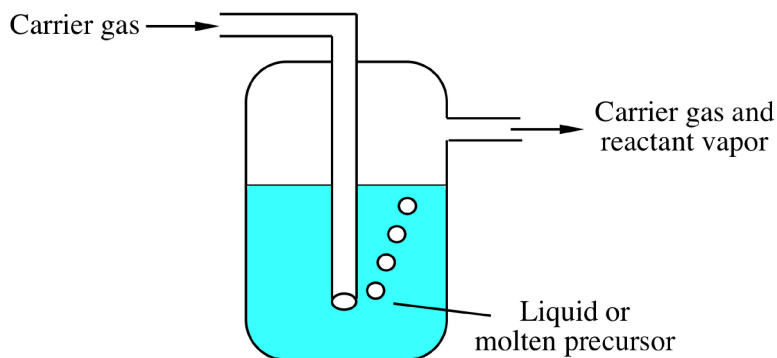
Demonstration of deposition rate on reactant concentration for CVD deposition. Adapted from J. G. Eden, in *Thin Film Processes II*, Eds. J. L. Vossen and W. Kern, Academic Press, New York (1991).

CVD systems

Precursor delivery

Flow of reactants into the reactor must be closely monitored to control stoichiometry and growth rate. Precursor delivery is very important since in many cases the flow rate can limit the deposition. For low vapor pressure solids, a carrier gas is passed over or through a bed of the heated solid to

transport the vapor into the reactor. Gas flow lines are usually heated to reduce condensation of the vapor in the flow lines. In the case of gas precursors, mass flowmeters easily gauge and control the flow rates. Liquid precursors are normally heated in a bubbler to achieve a desired vapor pressure ([link](#)).



Schematic representation of a bubbler for liquid precursors.

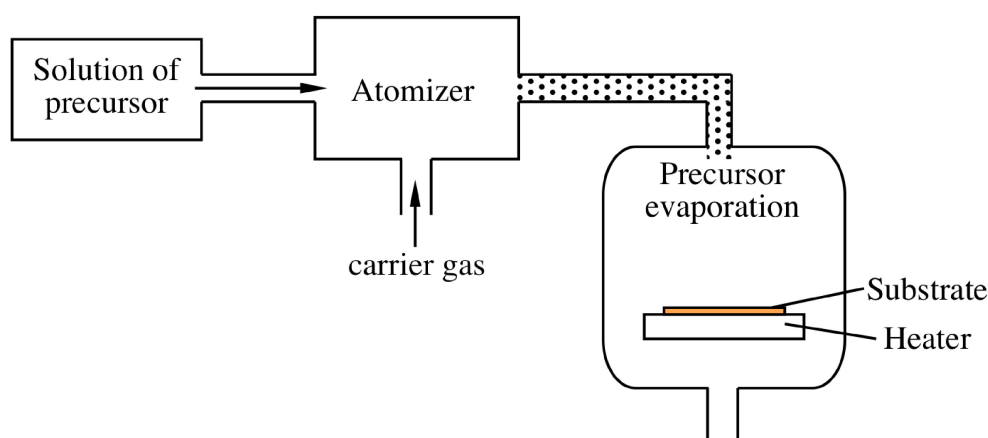
An inert gas such as hydrogen is bubbled through the liquid and by calculating the vapor pressure of the reactant and monitoring the flow rate of the hydrogen, the flow rate of the precursor is controlled by using [link](#), where Q_{MO} is the flow rate of the metal-organic precursor, Q_{H_2} is the flow rate of hydrogen through the bubbler, P_{MO} is the vapor pressure of the metal-organic at the bubbler temperature, and P_B is the pressure of the bubbler.

Equation:

$$Q_{MO} = Q_{H_2} \times \frac{P_{MO}}{P_B - P_{MO}}$$

Another method of introducing liquid precursors involves flash vaporization where the liquid is passed into a flask heated above the boiling

point of the liquid. The gas vapor is then passed through heated lines to the CVD chamber. Often, a carrier gas is added to provide a fixed flow rate into the reactor. This method of precursor introduction is useful when the precursor will decompose if heated over time. A similar technique called spray pyrolysis introduces the precursors in the form of aerosol droplets. The droplets evaporate in the chamber from the heated gas above the substrate or heated chamber walls ([\[link\]](#)). Then the reaction proceeds as a normal CVD process.

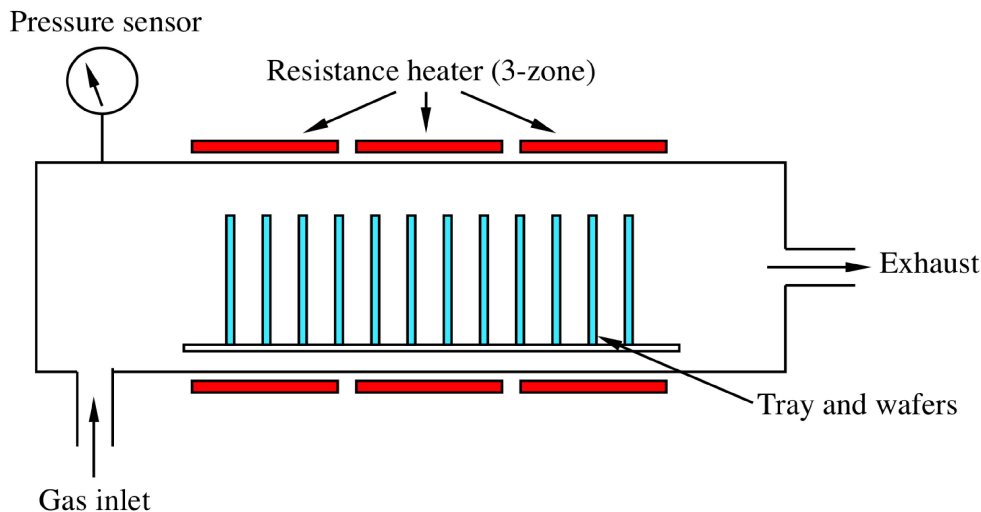


Schematic representation of a typical aerosol delivery system for CVD precursors. Adapted from T. T. Kodas and M. J. Hamton-Smith, *The Chemistry of Metal CVD*, VCH, New York (1994).

Thermal CVD reactors

In thermal CVD temperatures as high as 2000 °C may be needed to thermally decompose the precursors. Heating is normally accomplished by use of resistive heating, radio frequency (rf) induction heating, or radiant heating. There are two basic types of reactors for thermal CVD: the hot wall reactor and the cold wall reactor.

A hot wall reactor is an isothermal furnace into which the substrates are placed. Hot wall reactors are typically very large and depositions are done on several substrates at once. Since the whole chamber is heated, precise temperature control can be achieved with correct furnace design. A disadvantage of the hot wall configuration is that deposition occurs on the walls of the chamber as well as on the substrate. As a consequence, hot wall reactors must be frequently cleaned to reduce flaking of particles from the walls which may contaminate the substrates. Furthermore, reactions in the heated gas and at the walls deplete the reactants and can result in feed rate limited growth. [\[link\]](#) shows a typical low pressure hot wall CVD reactor.



Schematic of a typical low pressure hot wall CVD reactor used in coating silicon substrates. Adapted from H. O. Pierson, *Handbook of Chemical Vapor Deposition*, Noyes Publications, Park Ridge (1992).

In a cold wall reactor only the substrate is heated, usually by induction or radiant heating. Since most CVD reactions are endothermic, deposition is preferentially on the area of highest temperature. As a result, deposition is only on the substrate and the cooler reactor walls stay clean. Cold wall CVD has two main advantages over the hot wall configuration. First,

particulate contamination is reduced since there are no deposits formed on the walls of the reactor. Second, since decomposition only occurs on the substrate there is no depletion of source gases due to reaction on the walls. However, hot wall reactors tend to have higher throughput since the designs more easily accommodate multiple wafer configurations.

The final issue in design of a thermal CVD reactor is the operating pressure. The pressure of the reactor has a large effect on the rate limiting step of the deposition. Atmospheric pressure reactors have a large boundary layer ([\[link\]](#)) and non-uniform diffusion of reactants through the boundary layer often results in non-uniform film compositions across the wafer. Conversely, low pressure reactors have a nearly non-existent boundary layer and reactants easily diffuse to the substrate ([\[link\]](#)). However, the difficulty in maintaining a uniform temperature profile across the wafer can result in thickness non-uniformities since the deposition rate in low pressure reactors is strongly temperature dependent. Careful studies of the flow dynamics and temperature profiles of CVD reactors are always carried out in order to achieve uniform material depositions.

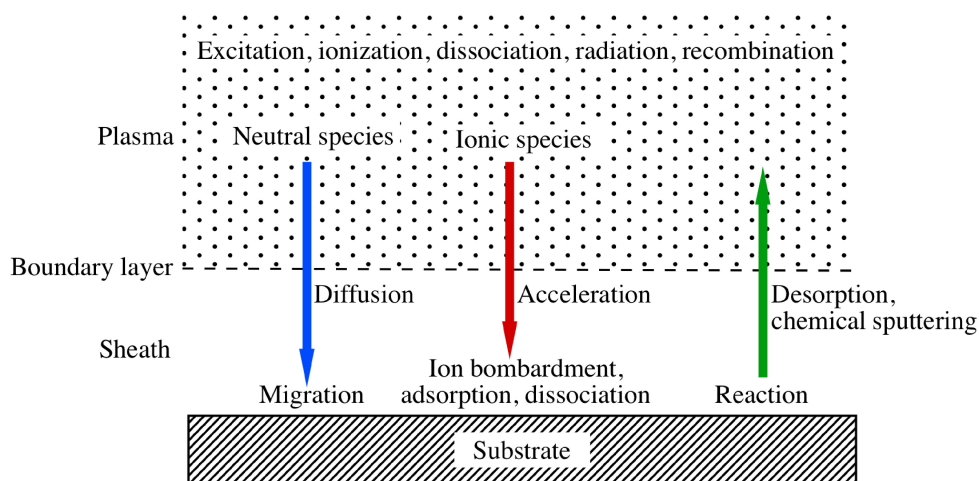
Plasma-enhanced CVD

Plasmas are generated for a variety of thin film processes including sputtering, etching, ashing, and plasma-enhanced CVD. Plasma-enhanced CVD (PECVD), sometimes called plasma-assisted (PACVD), has the advantage that plasma activated reactions occur at much lower temperatures compared to those in thermal CVD. For example, the thermal CVD of silicon nitride occurs between 700 - 900 °C, the equivalent PECVD process is accomplished between 250 - 350 °C.

A plasma is a partially ionized gas consisting of electrons and ions. Typical ionization fractions of 10^{-5} to 10^{-1} are encountered in process reactors. Plasmas are electrically conductive with the primary charge carriers being the electrons. The light mass of the electron allows it to respond much more quickly to changes in the field than the heavier ions. Most plasmas used for PECVD are generated using a rf electric field. In the high frequency electric field, the light electrons are quickly accelerated by the field but do not

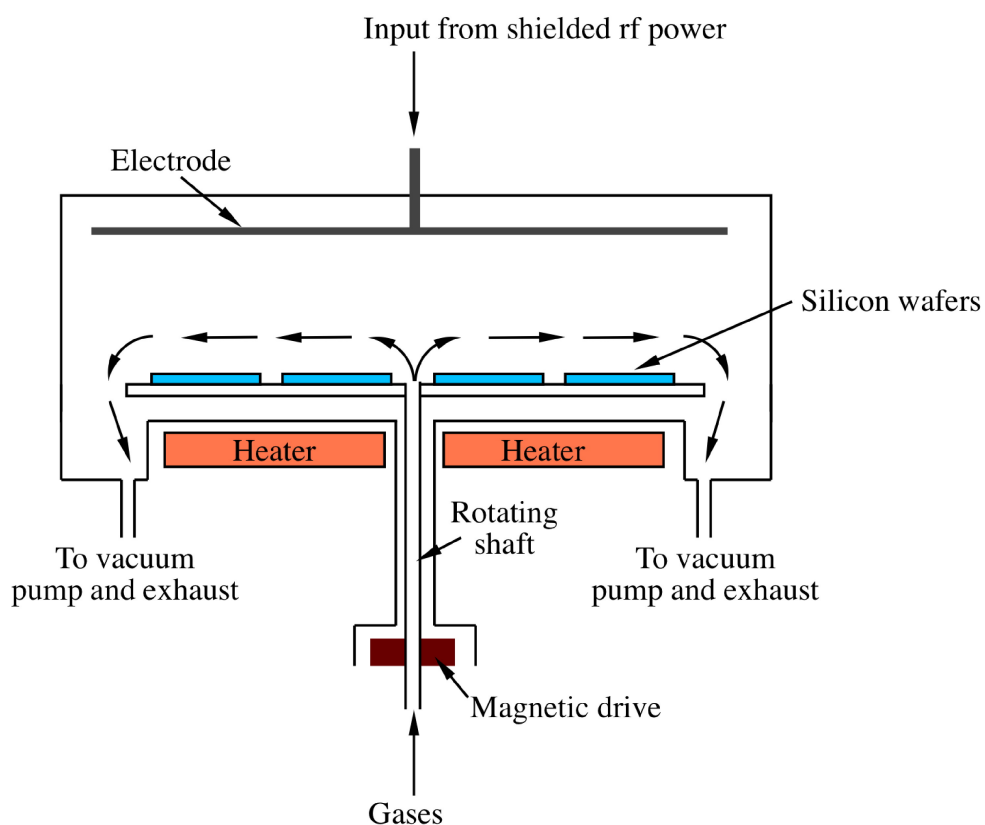
increase the temperature of the plasma because of their low mass. The heavy ions cannot respond to the quick changes in direction and therefore their temperature stays low. Electron energies in the plasma have a Maxwellian distribution in the 0.1 – 20 eV range. These energies are sufficiently high to excite molecules or break chemical bonds in collisions between electrons and gas molecules. The high energy electrons inelastically collide with gas molecules resulting in excitation or ionization. The reactive species generated by the collisions do not have the energy barriers to reactions that the parent precursors do. Therefore, the reactive ions are able to form films at temperatures much lower than those required for thermal CVD.

The general reaction sequence for PECVD is shown in [\[link\]](#). In addition to the processes that occur in thermal CVD, reactive species resulting from electron dissociation of parent molecules also diffuse to the surface. The reactive species have lower activation energies for chemical reactions and usually have higher sticking coefficients to the substrate. Therefore, the PECVD reaction is dominated by the reactive species on the surface and not any of the the parent precursor molecules that also diffuse to the surface.



Reaction sequence in PECVD. Adapted from M. Konuma, *Film Deposition by Plasma Techniques*, Springer-Verlag, New York (1992).

A basic PECVD reactor is shown in [\[link\]](#). The wafer chuck acts as the lower electrode and is normally placed at ground potential. Gases are either introduced radially at the edges of the reactor and pumped out from the center, or gases can be introduced from the center and pumped at the edges as shown in [\[link\]](#). The magnetic drive allows rotation of the wafers during processing to randomize substrate position. Some newer reactors introduce the gases through holes drilled in the upper electrode. This method of gas introduction gives a more uniform plasma distribution.



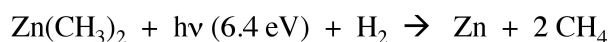
Schematic representation of a radial flow PECVD reactor. Adapted from H. O. Pierson, *Handbook of Chemical Vapor Deposition*, Noyes Publications, Park Ridge (1992).

Plasma CVD has numerous advantages over thermal CVD. Obviously the reduced deposition temperature is a bonus for the semiconductor industry which must worry about dopant diffusion and metal interconnects melting at the temperatures required for thermal CVD. Also, the low pressures (between 0.1 - 10 Torr) required for sustaining a plasma result in surface kinetics controlling the reaction and therefore greater film uniformity. A disadvantage of plasma CVD is that it is often difficult to control stoichiometry due to variations in bond strengths of various precursors. For example, PECVD films of silicon nitride tend to be silicon rich because of the relative bond strength of N₂ relative to the Si-H bond. Additionally, some films may be easily damaged by ion bombardment from the plasma.

Photochemical CVD

Photochemical CVD uses the energy of photons to initiate the chemical reactions. Photodissociation of the chemical precursor involves the absorption of one or more photons resulting in the breaking of a chemical bond. The most common precursors for photo-assisted deposition are the hydrides, carbonyls, and the alkyls. The dissociation of dimethylzinc by [\[link\]](#), a photon creates a zinc radical and a methyl radical ($\cdot\text{CH}_3$) that will react with hydrogen in the reactor to produce methane.

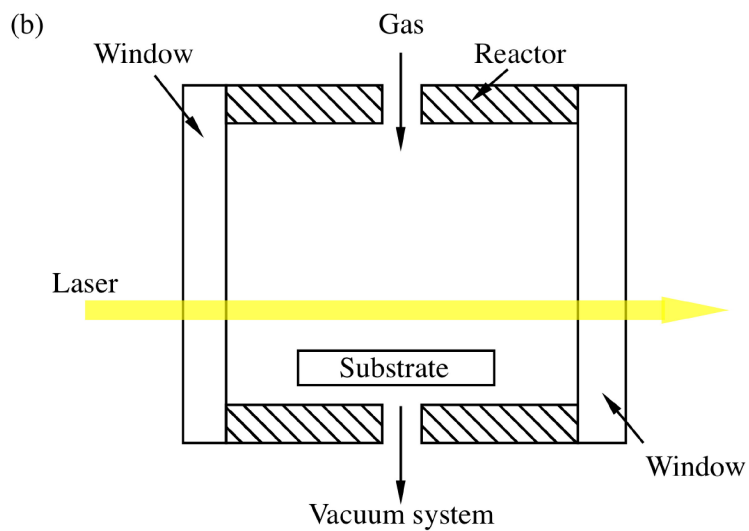
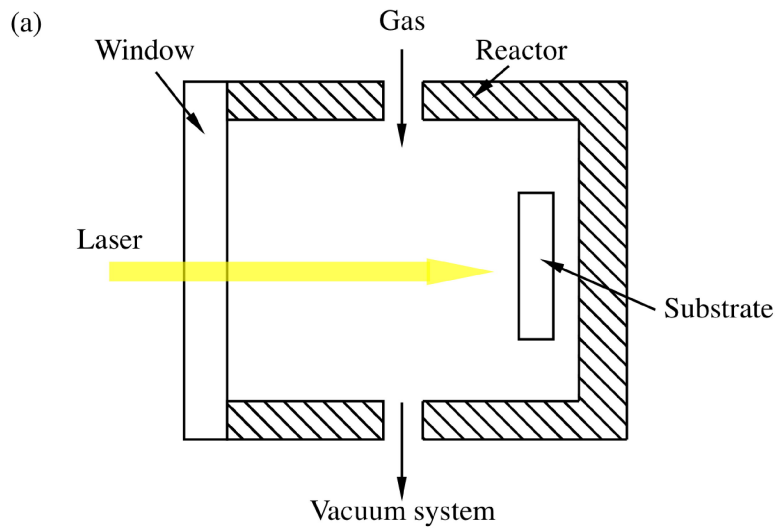
Equation:



Like several metal-organics, dimethylzinc is dissociated by the absorption of only one UV photon. However, some precursors require absorption of more than one photon to completely dissociate. There are two basic configurations for photochemical CVD. The first method uses a laser primarily as a localized heat source. The second method uses high energy photons to decompose the reactants on or near the growth surface.

In thermal laser CVD, sometimes referred to as laser pyrolysis, the laser is used to heat a substrate that absorbs the laser photons. Laser heating of substrates is a very localized process and deposition occurs selectively on the illuminated portions of the substrates. Except for the method of heating, laser CVD is identical to thermal CVD. The laser CVD method has the potential to be used for direct writing of features with relatively high resolution. The lateral extent of film growth when the substrate is illuminated by a laser is determined not only by the spot size of the laser, but by the thermal conductivity of the substrate. A variation of laser pyrolysis uses a laser to heat the gas molecules such that they are fragmented by thermal processes.

Photochemical effects can be induced by a laser if the precursor molecules absorb at the laser wavelength. UV photons have sufficient energy to break the bonds in the precursor chemicals. Laser-assisted CVD (LACVD) uses a laser, usually an excimer laser, to provide the high energy photons needed to break the bonds in the precursor molecules. [\[link\]](#) shows two geometries for LACVD. For the perpendicular illumination the photochemical effects generally occur in the adsorbed adlayer on the substrate. Perpendicular irradiation is often done using a UV lamp instead of a laser so that unwanted substrate heating is not produced by the light source. The parallel illumination configuration has the benefit that reaction by-products are produced further from the growth surface and have less chance of being incorporated into the growing film. The main benefit of LACVD is that nearly no heat is required for deposition of high quality films.



Parallel (a) and perpendicular (b) irradiation in laser CVD. Adapted from J. G. Eden, in *Thin Film Processes II*, Eds. J. L. Vossen and W. Kern, Academic Press, New York (1991).

An application of laser photolysis is photonucleation. Photonucleation is the process by which a chemisorbed adlayer of metal precursors is photolyzed

by the laser to create a nucleation site for further growth. Photonucleation is useful in promoting growth on substrates that have small sticking coefficients for gas phase metal atoms. By beginning the nucleation process with photonucleation the natural barrier to surface nucleation on the substrate is overcome.

Bibliography

- J. G. Eden, in *Thin Film Processes II*, Eds. J. L. Vossen and W. Kern, Academic Press, New York (1991).
- T. T. Kodas and M. J. Hamton-Smith, *The Chemistry of Metal CVD*, VCH, New York (1994).
- M. Konuma, *Film Deposition by Plasma Techniques*, Springer-Verlag, New York (1992).
- H. O. Pierson, *Handbook of Chemical Vapor Deposition*, Noyes Publications, Park Ridge (1992).
- R. Reif and W. Kern, in *Thin Film Processes II*, Eds. J. L. Vossen and W. Kern, Academic Press, New York (1991).
- G. B. Stringfellow, *Organometallic Vapor-Phase Epitaxy: Theory and Practice*, Academic Press, New York (1994).

Liquid Phase Deposition

Introduction

Silicon dioxide (silica, SiO_2) has been the most researched chemical compound apart from water. Silica has been used throughout history, for example, flint, which when sharpened formed one of humanities first tools. Crystalline silica, or sand, was melted into glass as early as 5000 B.C., birthing a technology that has gained sophistication in modern times. Silicon is the second most plentiful element in the Earth's crust, the most plentiful being oxygen. It is thus surprising that it was not until 1800 that silica was named a compound by Sir Humphry Davy. He, however, failed to isolate its components via electrolysis, and it is Jöns Jacob Berzelius who is thus credited with discovering silica in 1824. He heated potassium fluorosilicate with potassium metal and, after purifying the product of this reaction with water, produced amorphous silica powder.

The most common forms of silica employed in industry include α -quartz, vitreous silica, silica gel, fumed silica and diatomaceous earth. Synthetic quartz is hydrothermally grown from a seed crystal, with aqueous NaOH and vitreous SiO_2 , at 400 °C and 1.7 kbar. Because it is a piezoelectric material, it is used in crystal oscillators, transducers, pickups and filters for frequency control and modulation. Vitreous silica is super cooled liquid silica used in laboratory glassware, protective tubing sheaths and vapor grown films. Silica gel is formed from the reaction of aqueous sodium silicate with acid, after which it is washed and dehydrated. Silica gel is an exceptionally porous material with numerous applications including use as a dessicant, chromatographic support, catalyst substrate and insulator. Pyrogenic or fumed silica is produced by the high temperature hydrolysis, in an oxyhydrogen flame, of SiCl_4 . Its applications include use as a thickening agent and reinforcing filler in polymers. Diatomaceous earth, the ecto-skeletons of tiny unicellular marine algae called diatoms, is mined from vast deposits in Europe and North America. Its primary use is in filtration. Additional applications include use as an abrasive, insulator, filler and a lightweight aggregate.

Methods of colloidal growth and thin film deposition of amorphous silica have been investigated since 1925. The two most common and well-investigated methods of forming SiO_2 in a sol or as a film or coating are condensation of alkoxysilanes (known as the Stober method) and hydrolysis of metal alkoxides (the Iler or dense silica [DS] process).

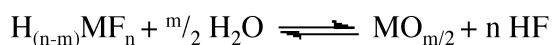
Liquid phase deposition (LPD)

LPD is a method for the “non-electrochemical production of polycrystalline ceramic films at low temperatures.” LPD, along with other aqueous solution methods [chemical bath deposition (CBD), successive ion layer adsorption and reaction (SILAR) and electroless deposition (ED) with catalyst] has developed as a potential substitute for vapor-phase and chemical-precursor systems. Aqueous solution methods are not dependent on vacuum systems or glove boxes, and the use of easily acquired reagents reduces reliance on expensive or sensitive organometallic precursors. Thus, LPD holds potential for reduced production costs and environmental impact. Films may be deposited on substrates that might not be chemically or mechanically stable at higher temperatures. In addition, the use of liquid as a deposition medium allows coating of non-planar substrates, expanding the range of substrates that are capable of being coated. Aqueous deposition techniques have not reached the level of maturation that vapor-phase techniques have in respect to a high level of control over composition, microstructure and growth rates of the resulting films, but their prospect makes them attractive for research.

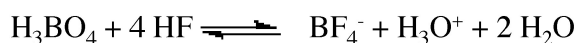
LPD generally refers to the formation of oxide thin films, the most common being SiO_2 , from an aqueous solution of a metal-fluoro complex $[\text{MF}_n]^{m-n}$, which is slowly hydrolyzed using water, boric acid or aluminum metal. Addition of water drives precipitation of the oxide. Boric acid and aluminum work as fluoride scavengers, rapidly weakening the fluoro complex and precipitating the oxide. These reactants are added either drop wise or outright, both methods allowing for high control of the hydrolysis reaction and of the solution’s supersaturation. Film formation is accomplished from highly acidic solutions, in contrast to the basic or weakly acidic solutions used in chemical bath deposition.

A generic description of the LPD reaction is shown in [\[link\]](#), where m is the charge on the metal cation. If the concentration of water is increased or the concentration of hydrofluoric acid (HF) is decreased, the equilibrium will be shifted toward the oxide. Use of boric acid or aluminum metal will accomplish the latter, see [\[link\]](#) and [\[link\]](#). The most popular of these methods for accomplishing oxide formation has been through the addition of boric acid.

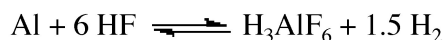
Equation:



Equation:



Equation:



The first patent using liquid phase deposition (LPD) of silicon dioxide via fluorosilicic acid solutions (H_2SiF_6) was granted to the Radio Corporation of America (RCA) in 1950. RCA used LPD as a method for coating anti-reflective films on glass, but the patent promised further applications. Since this initial patent there have been many further patents and papers utilizing this method, in variable forms, to coat substrates, usually silicon, with silicon dioxide. The impetus behind this work is to create an alternative to the growth of insulator coatings by thermal oxidation or chemical vapor deposition (CVD) for planar silicon chip technology. Thermal oxidation and CVD are performed at elevated temperatures, requiring a higher output of energy and more complicated instrumentation than that of LPD. The most simple and elegant of the LPD methods uses only water to catalyze silica thin film growth on silicon from a solution of fluorosilicic acid supersaturated with silicon dioxide, [\[link\]](#).

Equation:



The amount of water reacted with the supersaturated fluorosilicic acid solution controls both the growth rate and incorporation of fluorine into the resulting silica matrix. Both growth rate and fluorine content increase with increased addition of water. Ultimately this “dilution” affects the optical properties of the resulting silica film; an increased amount of fluorine decreases its dielectric constant (and thus its refractive index).

To ensure a uniform film growth with LPD, the preparation of the surface to be coated is of utmost importance. Suitable treatments may involve the formation of surface hydroxides, the pre-deposition or self-assembly of an appropriate seed layer. The most efficient coverage is seen with silicon surfaces functionalized with hydroxy (-OH) groups prior to immersion in the growth solution. This can be achieved through appropriate etching of the silicon surface. It is proposed that the silanol (Si-OH) groups act to seed the growth of the silica film through condensation reactions with the silicic acid formed in the growth solution.

Lee and co-workers and Homma separately propose that intermediate, hydrolyzed species, $\text{SiF}_n(\text{OH})_{4-n}$ ($n < 4$), are formed by the reaction shown in [\[link\]](#). According to Lee, these species then react with the substrate surface to form a film. Homma proposes that fluorine-containing siloxanes are subsequently formed, which adsorb onto the surface where condensation and bonding occurs between the oligomers and surface hydroxyl groups. The former mechanism implies a molecular growth mechanism, whereas the latter implies homogeneous nucleation with subsequent deposition.

Equation:



In concentrated fluorosilicic acid solutions silica can be dissolved to well beyond its solubility, forming fluorosilicon complexes such as $[\text{SiF}_6.\text{SiF}_4]^{2-}$, [\[link\]](#). The bridged fluorosilicon complex has electron deficient silicon

because of the high electronegativity of the bonded fluorines, creating weak Si-F bonds. These bonds are then prone to nucleophilic attack by water. The fluorine ion (F^-) combines with the proton in this reaction to form hydrofluoric acid (HF). The product of this reaction can then react further with water to yield $[SiF_4(OH)_2]^{2-}$, SiF_4 and HF. The high acidity of the solution then allows protons to react with $[SiF_4(OH)_2]^{2-}$ to form tetrafluorosilicate (SiF_4) and water, [\[link\]](#). Hydrolysis of the SiF_4 will then yield the hexafluorosilicate anion, protons and silicic acid, [\[link\]](#).

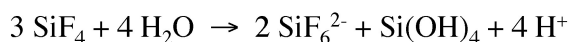
Equation:



Equation:

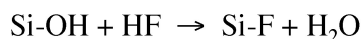


Equation:



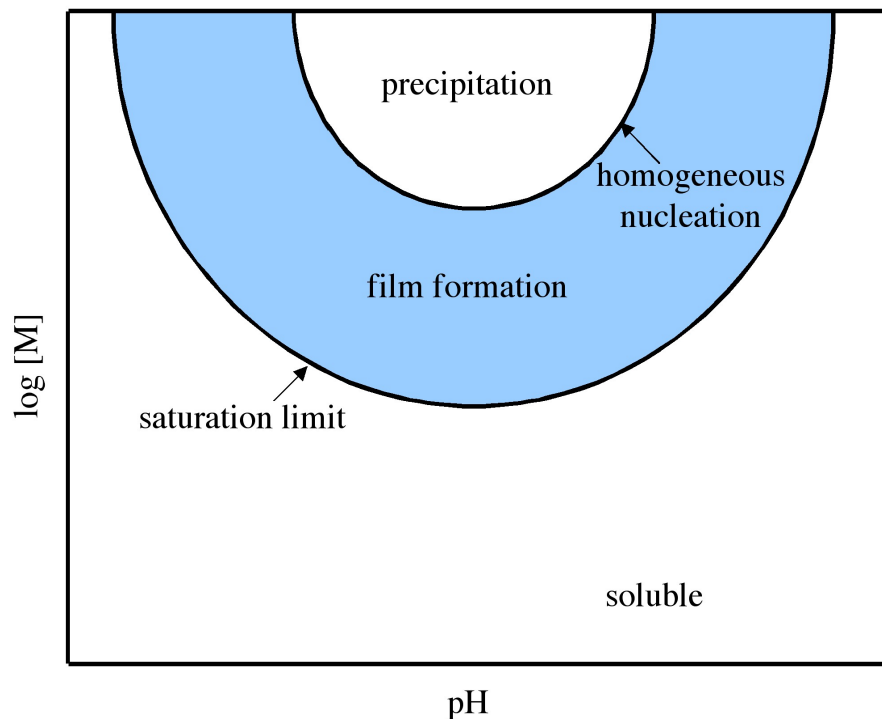
Silicic acid is adsorbed onto the surface of the substrate that has been introduced into the growth solution. Molecular growth of silica on the substrate surface is initialized in an acid catalyzed dehydration between the silicic acid and the silanol groups on the substrate surface. Si-O-Si bonds are formed, resulting in an initial silica coating of the surface. Following reactions between the initial silica coating and the monosilicic acid in solution result in further silica deposition and growth. Because of the presence of HF in the solution, the surface and growing silica matrix is subject to attack according to the reaction in [\[link\]](#). This explains the incorporation of a quantity of fluorine into the silica film. Additionally, it reveals that a certain amount of silica etching occurs along with growth. Because of the prevalence of the silicic acid in the solution, however, deposition is predominant.

Equation:



This proposed mechanism, which is more in depth than those proposed by Lee and Homma, elucidates what is experimentally proven. The deposition rate of the silica increases with addition of H₂O because the nucleophilic attack of the fluorosilicon complex is then augmented, increasing the concentration of silicic acid in the growth solution. The H₂O addition increases the reaction rate and thus the concentration of HF in the growth solution, resulting in greater incorporation of fluorine into the silica matrix because of HF attack of the deposited film. Additionally, Yeh's mechanism supports a molecular growth model, i.e., heterogeneous growth, which represents a consensus of the body of research performed thus far.

In a solution with dissolved ceramic precursors, nucleation and growth will occur either in solution (homogenous nucleation) or on the surfaces of introduced solid phases (heterogeneous nucleation). Successful film formation relies on the promotion of heterogeneous nucleation. Solubility generally depends on the solution pH and the concentration of the species in solution. As the solution crosses over from a solvated state to a state of supersaturation, film formation can occur. It is vital to assure that the state of supersaturation is one that promotes film growth and not homogeneous nucleation and precipitation. This concept is illustrated in [\[link\]](#).



Idealized solubility diagram for film forming species in water. Adapted from B. C. Bunker, P. C. Rieke, B. J. Tarasevich, A. A. Campbell, G. E. Fryxall, G. L. Graff, L. Song, J. Liu, J. W. Virden, and G. L. McVay, *Science*, 1994, **264**, 48.

Silica can be dissolved in fluorosilicic acid to well above its solubility in water, which is approximately 220 ppm (mg/L). Depending on the concentration of the fluorosilicic acid solution, it can contain up to 20% more silica than is implied by the formula H_2SiF_6 . After saturation of the solution with SiO_2 , the solvated species is a mixture of fluorosilicates, which reacts as explained earlier. It must be emphasized that addition of water in this reaction is not simply dilution, but is the addition of a reactant, which places the solution in a metastable state (the blue area in [\[link\]](#)) in preparation for the introduction of a suitable surface to seed the growth of silica.

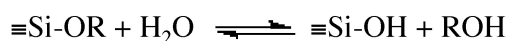
Another important factor in solution growth methods is interfacial energy. When a substrate with lower interfacial energy than that of a growing homogeneous nucleus is introduced into a growth solution, heterogeneous growth is favored. Thus, a seeded growth mechanism by definition introduces a substrate of lower interfacial energy into a supersaturated solution, facilitating heterogeneous growth. Lower interfacial energies can be a product of surface modification, as well as a property of the materials' natural state.

Comparing LPD to sol-gel

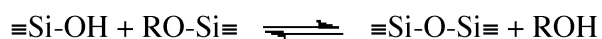
An alternative method to LPD for forming silica thin films is the sol-gel method. A sol is a colloidal dispersion of particles in a liquid. A gel is a material that contains a continuous solid matrix enclosing a continuous liquid phase. The liquid inhibits the solid from collapsing and the solid impedes release of the liquid. A formal definition of sol-gel processing is the “growth of colloidal particles and their linking together to form a gel.” This method describes both the hydrolysis and condensation of silicon alkoxides and the hydrolysis and condensation of aqueous silicates (the DS process).

In the hydrolysis of silicon alkoxides, an alkoxide group is replaced with a hydroxyl group, [\[link\]](#). Further condensation reactions between alkoxy groups or hydroxyl groups produce siloxane bonds, see [\[link\]](#) and [\[link\]](#).

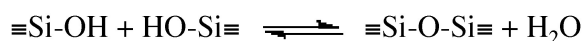
Equation:



Equation:



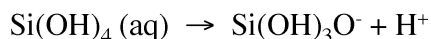
Equation:



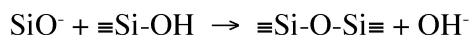
Tetramethoxysilane [Si(OMe)₄, TMOS] and tetraethylorthoxysilane [Si(OEt)₄, TEOS] are the most commonly used precursors in silica sol-gel processing. The alkoxides are hydrolyzed in their parent alcohols, with a mineral acid or base catalyst, producing silicate gels that can be deposited as coatings. The Stober method, which utilizes this chemistry, relies on homogeneous nucleation to produce monodisperse sols.

Iler's DS method of silica film formation was originally patented as a pigment coating to increase dispersibility of titania particles for use in the paint industry. The DS method is based on the aqueous chemistry of silica and takes advantage of the species present in solution at varying pH. Below pH 7 three-dimensional gel networks are formed. Above pH 7 silica surfaces are quite negatively charged ([\[link\]](#)), so that particle growth occurs without aggregation. The isoelectric point of silica is pH 2. Reactions above and below pH 2 are thought to occur through bimolecular nucleophilic condensation mechanisms. Above pH 2 an anionic species attacks a neutral species ([\[link\]](#)) and below pH 2 condensation involves a protonated silanol ([\[link\]](#)). The DS process has been utilized extensively in sol-gel coating technology and as a growth method for monodisperse and polydisperse sols.

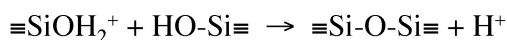
Equation:



Equation:



Equation:



Bibliography

- B. C. Bunker, P. C. Rieke, B. J. Tarasevich, A. A. Campbell, G. E. Fryxall, G. L. Graff, L. Song, J. Liu, J. W. Virden, and G. L. McVay,

Science, 1994, **264**, 48.

- P.-H. Chang, C.-T. Huang, and J.-S. Shie, *J. Electrochem. Soc.*, 1997, **144**, 1144.
- J.-S. Chou and S.-C. Lee, *J. Electrochem. Soc.*, 1994, **141**, 3214.
- T. Homma, T. Katoh, Y. Yamada, and Y. Murao, *J. Electrochem. Soc.*, 1993, **140**, 2410.
- R. K. Iler, *The Chemistry of Silica Solubility, Polymerization, Colloid and Surface Properties, and Biochemistry*, John Wiley & Sons (1979).
- H. R. Jafry, E. A. Whitsitt, and A. R. Barron, *J. Mater. Sci.*, 2007, **42**, 7381.
- T. Niesen and M. R. De Guire, *J. Electroceramics*, 2001, **6**, 169.
- N. Ozawa, Y. Kumazawa, and T. Yao, *Thin Solid Films*, 2002, **418**, 102.
- W. Stober, A. Fink, and E. Bohn, *J. Colloid Interface Sci.*, 1968, **26**, 62.
- D. Whitehouse, *Glass of the Roman Empire*, Corning (1988).
- E. A. Whitsitt and A. R. Barron, *Nano Lett.*, 2003, **3**, 775.
- E. A. Whitsitt and A. R. Barron, *Chem. Commun.*, 2003, 1042.
- E. A. Whitsitt and A. R. Barron, *J. Colloid Interface Sci.*, 2005, **287**, 318.
- C.-F. Yeh, C.-L. Chen, and G.-H. Lin, *J. Electrochem. Soc.*, 1994, **141**, 3177.

Selecting a Molecular Precursor for Chemical Vapor Deposition

Introduction

The proven utility of chemical vapor deposition (CVD) in a wide range of electronic materials systems (semiconductors, conductors, and insulators) has driven research efforts to investigate the potential for thin film growth of other materials, including: high temperature superconducting metal oxides, piezoelectric material, etc. Moreover, CVD potentially is well suited for the preparation of thin films on a wide range of substrates, including those of nonplanar geometries. CVD offers the advantages of mild process conditions (i.e., low temperatures), control over microstructure and composition, high deposition rates, and possible large scale processing. As with any CVD process, however, the critical factor in the deposition process has been the selection of precursors with suitable transport properties.

Factors in selecting a CVD precursor molecule

The following properties are among those that must be considered when selecting suitable candidates for a CVD precursor:

1. The precursor should be either a liquid or a solid, with sufficient vapor pressure and mass transport at the desired temperature, preferably below 200 °C. Liquids are preferred over solids, due to the difficulty of maintaining a constant flux of source vapors over a non-equilibrium percolation (solid) process. Such non-bubbling processes are a function of surface area, a non-constant variable with respect both to time and particle size. The upper temperature limit is not dictated by chemical factors; rather, it is a limitation imposed by the stability of the mass flow controllers and pneumatic valves utilized in commercial deposition equipment. It must be stressed that while the achievement of an *optimum* vapor pressure for efficient utilization as an industrially practicable source providing high film growth rates (>10 Torr at 25 °C) is a worthy goal, the usable pressure regimes are those in which evaluation can be carried out on compounds which exhibit vapor pressures exceeding 1 Torr at 100 °C.

2. The precursor must be chemically and thermally stable in the region bordered by the evaporation and transport temperatures, even after prolonged use. Early workers were plagued by irreproducible film growth results caused by premature decomposition of source compounds in the bubbler, in transfer lines, and, basically everywhere *except* on the substrate. Such experiences are to be avoided!
3. By its very nature, CVD demands a decomposable precursor. This generally is accomplished thermally; however, the plasma-enhanced growth regime has seen much improvement. In addition, photolytic processes have tremendous potential. Nevertheless, the precursor must be thermally robust *until deposition conditions are employed*.
4. The precursor should be relatively easy to synthesize, ensuring sufficient availability of material for testing and fabrication. It also is important that the synthesis of the compound be reproducible. It should be simple to prepare and purify to a relatively high level of purity. It should be non-toxic and *environmentally friendly* (i.e., as low a toxicity as can be attained, given the fundamental toxicity of particular elements such as mercury, thallium, barium, etc.). It should be routine to reproduce and scale-up the preparation for further developmental studies. It should utilize readily available starting reagents, and proceed by a minimum number of chemical transformations in order to minimize the cost.
5. Due to handling considerations, the source should be oxidatively, hydrolytically, thermally and photochemically stable under normal storage conditions, in addition the precursor should resist oligomerization (in the solid, liquid, or gaseous states). It is worth noting that practitioners of metal organic CVD (MOCVD), especially for 13-15 materials have of necessity become expert in the handling of very toxic, highly air sensitive materials.

Historically, researchers were limited in their choices of precursors to those that were readily known and commercially available. It must be emphasized that *none* of these previously known compounds had been designed specifically to serve as vapor phase transport molecules for the associated element. Thus, the scope was often limited to what was commercially available. However, as new compounds have now been made with the

specific goal of providing ideal CVD precursors the choice to academia and industry has increased.

Bibliography

- G. B. Stringfellow, *Organometallic Vapor Phase Epitaxy: Theory and Practice*, Academic Press, New York (1989).

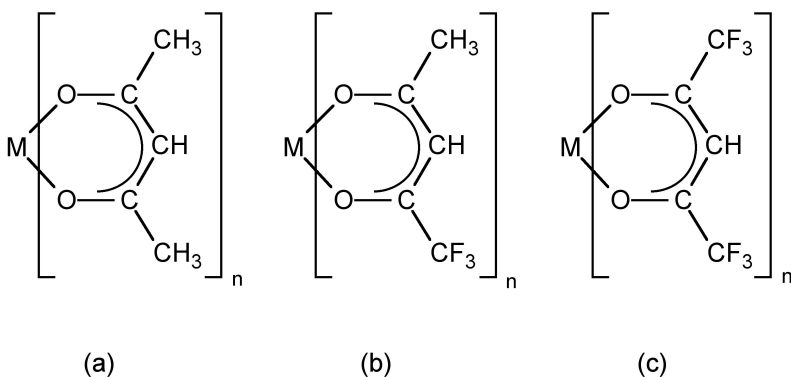
Determination of Sublimation Enthalpy and Vapor Pressure for Inorganic and Metal-Organic Compounds by Thermogravimetric Analysis

Introduction

Metal compounds and complexes are invaluable precursors for the chemical vapor deposition (CVD) of metal and non-metal thin films. In general, the precursor compounds are chosen on the basis of their relative volatility and their ability to decompose to the desired material under a suitable temperature regime. Unfortunately, many readily obtainable (commercially available) compounds are not of sufficient volatility to make them suitable for CVD applications. Thus, a *prediction* of the volatility of a metal-organic compounds as a function of its ligand identity and molecular structure would be desirable in order to determine the suitability of such compounds as CVD precursors. Equally important would be a method to determine the vapor pressure of a potential CVD precursor as well as its optimum temperature of sublimation.

It has been observed that for organic compounds it was determined that a rough proportionality exists between a compound's melting point and sublimation enthalpy; however, significant deviation is observed for inorganic compounds.

Enthalpies of sublimation for metal-organic compounds have been previously determined through a variety of methods, most commonly from vapor pressure measurements using complex experimental systems such as Knudsen effusion, temperature drop microcalorimetry and, more recently, differential scanning calorimetry (DSC). However, the measured values are highly dependent on the experimental procedure utilized. For example, the reported sublimation enthalpy of $\text{Al}(\text{acac})_3$ ([link](#)), where $M = \text{Al}$, $n = 3$) varies from 47.3 to 126 kJ/mol.



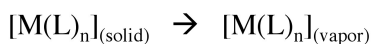
Structure of a typical metal β -diketonate complex. (a) acetylacetonate (acac); (b) trifluoro acetylacetonate (tfac), and (c) hexafluoroacetylacetonate (hfac).

Thermogravimetric analysis offers a simple and reproducible method for the determination of the vapor pressure of a potential CVD precursor as well as its enthalpy of sublimation.

Determination of sublimation enthalpy

The enthalpy of sublimation is a quantitative measure of the volatility of a particular solid. This information is useful when considering the feasibility of a particular precursor for CVD applications. An ideal sublimation process involves no compound decomposition and only results in a solid-gas phase change, i.e., [\[link\]](#).

Equation:



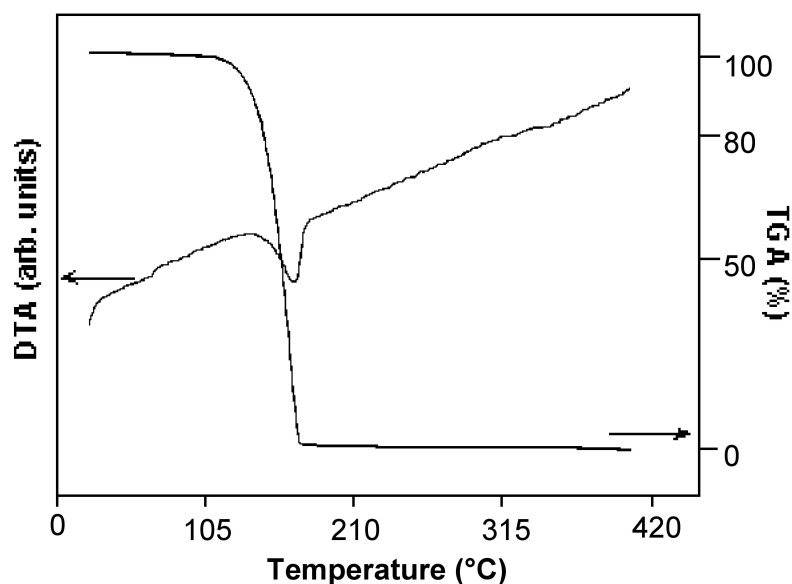
Since phase changes are thermodynamic processes following zero-order kinetics, the evaporation rate or rate of mass loss by sublimation (m_{sub}), at a constant temperature (T), is constant at a given temperature, [\[link\]](#).

Therefore, the m_{sub} values may be directly determined from the linear mass loss of the TGA data in isothermal regions.

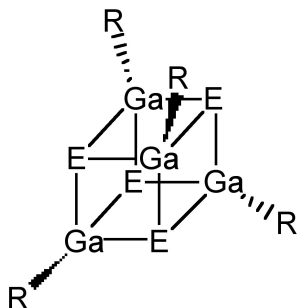
Equation:

$$m_{\text{sub}} = \frac{\Delta[\text{mass}]}{\Delta t}$$

The thermogravimetric and differential thermal analysis of the compound under study is performed to determine the temperature of sublimation and thermal events such as melting. [\[link\]](#) shows a typical TG/DTA plot for a gallium chalcogenide cubane compound ([\[link\]](#)).



A typical thermogravimetric/differential thermal analysis (TG/DTA) analysis of $[(\text{EtMe}_2\text{C})\text{GaSe}]_4$, whose structure is shown in [\[link\]](#). Adapted from E. G. Gillan, S. G. Bott, and A. R. Barron, *Chem. Mater.*, 1997, **9**, 3, 796.



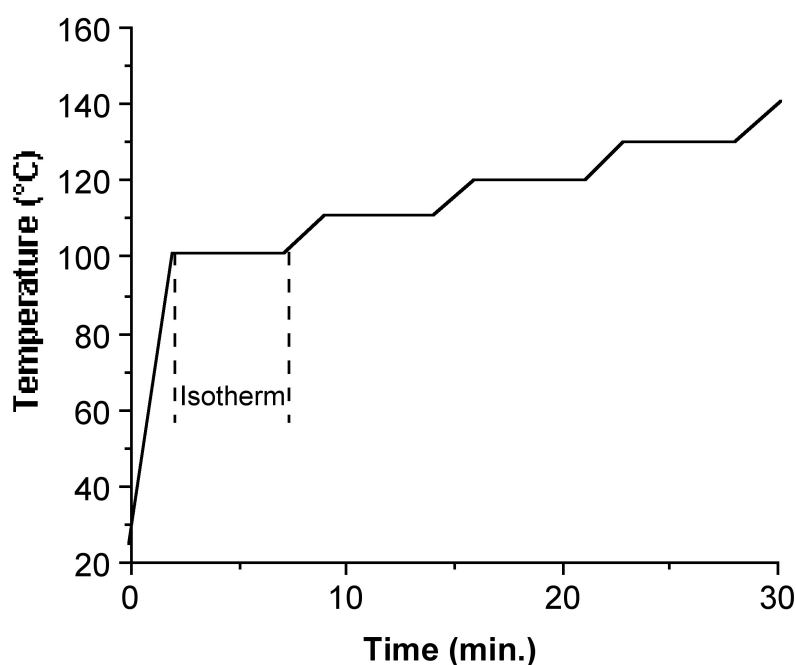
Structure of
gallium
chalcogenide
cubane
compound,
where E = S,
Se, and R =
CMe₃,
CMe₂Et,
CEt₂Me,
CEt₃.

Data collection

In a typical experiment 5 - 10 mg of sample is used with a heating rate of ca. 5 °C/min up to under either a 200-300 mL/min inert (N₂ or Ar) gas flow or a dynamic vacuum (*ca.* 0.2 Torr if using a typical vacuum pump). The argon flow rate was set to 90.0 mL/min and was carefully monitored to ensure a steady flow rate during runs and an identical flow rate from one set of data to the next.

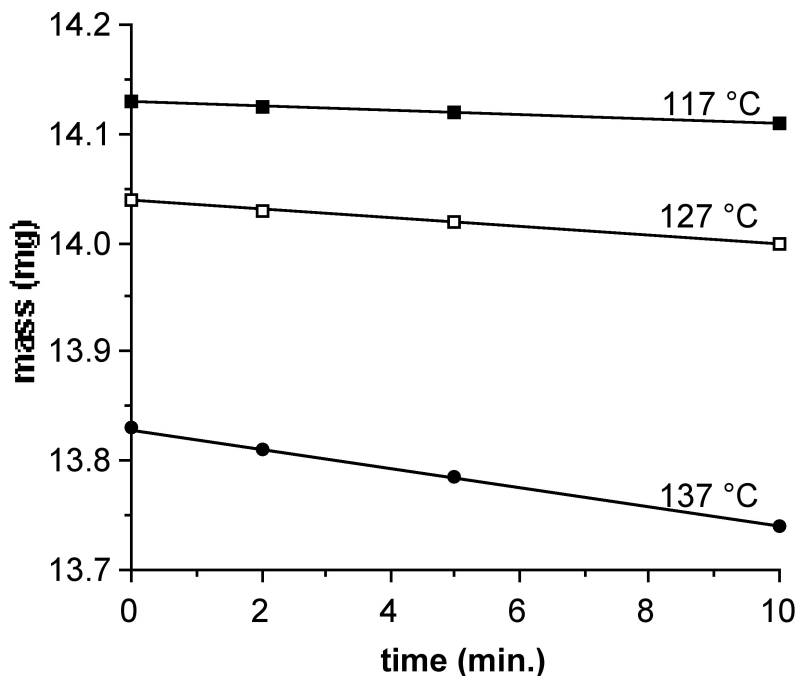
Once the temperature range is defined, the TGA is run with a preprogrammed temperature profile ([\[link\]](#)). It has been found that sufficient data can be obtained if each isothermal mass loss is monitored over a period (between 7 and 10 minutes is found to be sufficient) before

moving to the next temperature plateau. In all cases it is important to confirm that the mass loss at a given temperature is linear. If it is not, this can be due to either (a) temperature stabilization had not occurred and so longer times should be spent at each isotherm, or (b) decomposition is occurring along with sublimation, and lower temperature ranges must be used. The slope of each mass drop is measured and used to calculate sublimation enthalpies as discussed below.



A typical temperature profile for determination of isothermal mass loss rate.

As an illustrative example, [\[link\]](#) displays the data for the mass loss of $\text{Cr}(\text{acac})_3$ ([\[link\]](#)a, where $M = \text{Cr}$, $n = 3$) at three isothermal regions under a constant argon flow. Each isothermal data set should exhibit a linear relation. As expected for an endothermal phase change, the linear slope, equal to m_{sub} , increases with increasing temperature.



Plot of TGA results for Cr(acac)₃ performed at different isothermal regions. Adapted from B. D. Fahlman and A. R. Barron, *Adv. Mater. Optics Electron.*, 2000, **10**, 223.

Note: Samples of iron acetylacetonate ([link](#)), where M = Fe, n = 3) may be used as a calibration standard through ΔH_{sub} determinations before each day of use. If the measured value of the sublimation enthalpy for Fe(acac)₃ is found to differ from the literature value by more than 5%, the sample is re-analyzed and the flow rates are optimized until an appropriate value is obtained. Only after such a calibration is optimized should other complexes be analyzed. It is important to note that while small amounts (< 10%) of involatile impurities will not interfere with the ΔH_{sub} analysis, competitively volatile impurities will produce higher apparent sublimation rates.

It is important to discuss at this point the various factors that must be controlled in order to obtain meaningful (useful) m_{sub} data from TGA data.

1. The sublimation rate is independent of the amount of material used but may exhibit some dependence on the flow rate of an inert carrier gas, since this will affect the equilibrium concentration of the cubane in the vapor phase. While little variation was observed we decided that for consistency m_{sub} values should be derived from vacuum experiments only.
2. The surface area of the solid in a given experiment should remain approximately constant; otherwise the sublimation rate (i.e., mass/time) at different temperatures cannot be compared, since as the relative surface area of a given crystallite decreases during the experiment the apparent sublimation rate will also decrease. To minimize this problem, data was taken over a small temperature ranges (*ca.* 30 °C), and overall sublimation was kept low (*ca.* 25% mass loss representing a surface area change of less than 15%). In experiments where significant surface area changes occurred the values of m_{sub} deviated significantly from linearity on a $\log(m_{\text{sub}})$ versus $1/T$ plot.
3. The compound being analyzed must not decompose to any significant degree, because the mass changes due to decomposition will cause a reduction in the apparent m_{sub} value, producing erroneous results. With a simultaneous TG/DTA system it is possible to observe exothermic events if decomposition occurs, however the clearest indication is shown by the mass loss versus time curves which are no longer linear but exhibit exponential decays characteristic of first or second order decomposition processes.

Data analysis

The basis of analyzing isothermal TGA data involves using the Clausius-Clapeyron relation between vapor pressure (p) and temperature (T), [\[link\]](#), where ΔH_{sub} is the enthalpy of sublimation and R is the gas constant (8.314 J/K.mol).

Equation:

$$\frac{d \ln(p)}{dT} = \frac{\Delta H_{\text{sub}}}{RT^2}$$

Since m_{sub} data are obtained from TGA data, it is necessary to utilize the Langmuir equation, [\[link\]](#), that relates the vapor pressure of a solid with its sublimation rate.

Equation:

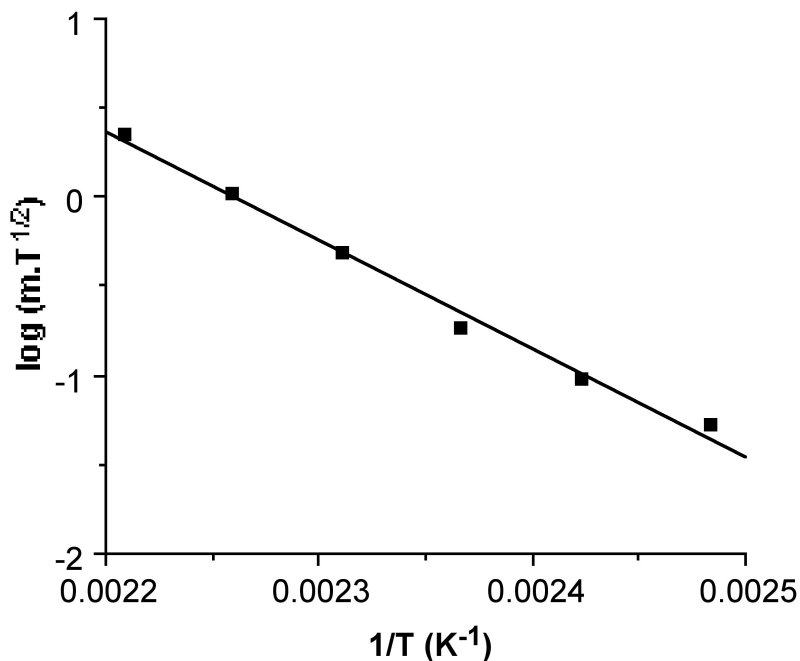
$$p = \left[\frac{2\pi RT}{M_w} \right]^{0.5} m_{\text{sub}}$$

After integrating [\[link\]](#) in log form, substituting in [\[link\]](#), and consolidating the constants, one obtains the useful equality, [\[link\]](#).

Equation:

$$\log(m_{\text{sub}}\sqrt{T}) = \frac{-0.0522(\Delta H_{\text{sub}})}{T} + \left[\frac{0.0522(\Delta H_{\text{sub}})}{T_{\text{sub}}} - \frac{1}{2} \log\left(\frac{1306}{M_w}\right) \right]$$

Hence, the linear slope of a $\log(m_{\text{sub}}T^{1/2})$ versus $1/T$ plot yields ΔH_{sub} . An example of a typical plot and the corresponding ΔH_{sub} value is shown in [\[link\]](#). In addition, the y intercept of such a plot provides a value for T_{sub} , the calculated sublimation temperature at atmospheric pressure.



Plot of $\log(m_{\text{sub}} T^{1/2})$ versus $1/T$ and the determination of the ΔH_{sub} (112.6 kJ/mol) for $\text{Fe}(\text{acac})_3$ ($R^2 = 0.9989$). Adapted from B. D. Fahlman and A. R. Barron, *Adv. Mater. Optics Electron.*, 2000, **10**, 223.

[\[link\]](#) lists the typical results using the TGA method for a variety of metal β -diketonates, while [\[link\]](#) lists similar values obtained for gallium chalcogenide cubane compounds.

Compound	ΔH_{sub} (kJ/mol)	ΔS_{sub} (J/K.mol)	T_{sub} calc. (°C)	Calculated vapor pressure @ 150 °C (Torr)

Al(acac) ₃	93	220	150	3.261
Al(tfac) ₃	74	192	111	9.715
Al(hfac) ₃	52	152	70	29.120
Cr(acac) ₃	91	216	148	3.328
Cr(tfac) ₃	71	186	109	9.910
Cr(hfac) ₃	46	134	69	29.511
Fe(acac) ₃	112	259	161	2.781
Fe(tfac) ₃	96	243	121	8.340
Fe(hfac) ₃	60	169	81	25.021
Co(acac) ₃	138	311	170	1.059
Co(tfac) ₃	119	295	131	3.319
Co(hfac) ₃	73	200	90	9.132

Selected thermodynamic data for metal β -diketonate compounds determined from thermogravimetric analysis. Data from B. D. Fahlman and A. R. Barron, *Adv. Mater. Optics Electron.*, 2000, **10**, 223.

Compound	ΔH_{sub} (kJ/mol)	ΔS_{sub} (J/K. mol)	T _{sub} calc. (°C)	Calculated vapor pressure
----------	-------------------------------------	--	-----------------------------------	---------------------------------

				@ 150 °C (Torr)
$[(\text{Me}_3\text{C})\text{GaS}]_4$	110	300	94	22.75
$[(\text{EtMe}_2\text{C})\text{GaS}]_4$	124	330	102	18.89
$[(\text{Et}_2\text{MeC})\text{GaS}]_4$	137	339	131	1.173
$[(\text{Et}_3\text{C})\text{GaS}]_4$	149	333	175	0.018
$[(\text{Me}_3\text{C})\text{GaSe}]_4$	119	305	116	3.668
$[(\text{EtMe}_2\text{C})\text{GaSe}]_4$	137	344	124	2.562
$[(\text{Et}_2\text{MeC})\text{GaSe}]_4$	147	359	136	0.815
$[(\text{Et}_3\text{C})\text{GaSe}]_4$	156	339	189	0.005

Selected thermodynamic data for gallium chalcogenide cubane compounds determined from thermogravimetric analysis. Data from E. G. Gillan, S. G. Bott, and A. R. Barron, *Chem. Mater.*, 1997, **9**, 3, 796.

A common method used to enhance precursor volatility and corresponding efficacy for CVD applications is to incorporate partially ([link](#)b) or fully ([link](#)c) fluorinated ligands. As may be seen from [link](#) this substitution does results in significant decrease in the ΔH_{sub} , and thus increased volatility. The observed enhancement in volatility may be rationalized either by an increased amount of intermolecular repulsion due to the additional lone pairs or that the reduced polarizability of fluorine (relative to hydrogen) causes fluorinated ligands to have less intermolecular attractive interactions.

Determination of sublimation entropy

The entropy of sublimation is readily calculated from the ΔH_{sub} and the calculated T_{sub} data, [link](#).

Equation:

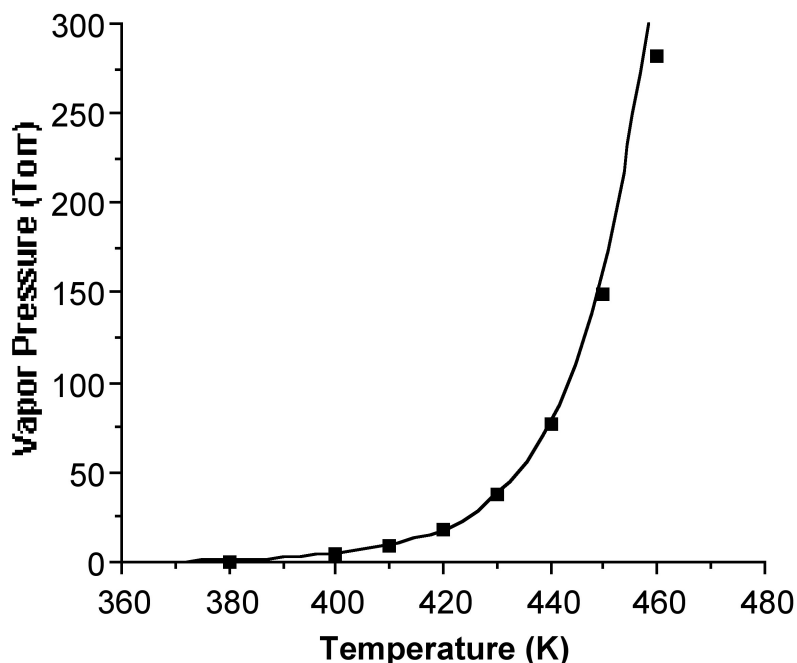
$$\Delta S_{\text{sub}} = \frac{\Delta H_{\text{sub}}}{T_{\text{sub}}}$$

[\[link\]](#) and [\[link\]](#) show typical values for metal β -diketonate compounds and gallium chalcogenide cubane compounds, respectively. The range observed for gallium chalcogenide cubane compounds ($\Delta S_{\text{sub}} = 330 \pm 20$ J/K.mol) is slightly larger than values reported for the metal β -diketonates compounds ($\Delta S_{\text{sub}} = 130 - 330$ J/K.mol) and organic compounds (100 - 200 J/K.mol), as would be expected for a transformation giving translational and internal degrees of freedom. For any particular chalcogenide, i.e., $[(R)\text{GaS}]_4$, the lowest ΔS_{sub} are observed for the Me_3C derivatives, and the largest ΔS_{sub} for the Et_2MeC derivatives, see [\[link\]](#). This is in line with the relative increase in the modes of freedom for the alkyl groups in the absence of crystal packing forces.

Determination of vapor pressure

While the sublimation temperature is an important parameter to determine the suitability of a potential precursor compounds for CVD, it is often preferable to express a compound's volatility in terms of its vapor pressure. However, while it is relatively straightforward to determine the vapor pressure of a liquid or gas, measurements of solids are difficult (e.g., use of the isoteniscopic method) and few laboratories are equipped to perform such experiments. Given that TGA apparatus are increasingly accessible, it would therefore be desirable to have a simple method for vapor pressure determination that can be accomplished on a TGA.

Substitution of [\[link\]](#) into [\[link\]](#) allows for the calculation of the vapor pressure (p) as a function of temperature (T). For example, [\[link\]](#) shows the calculated temperature dependence of the vapor pressure for $[(\text{Me}_3\text{C})\text{GaS}]_4$. The calculated vapor pressures at 150 °C for metal β -diketonates compounds and gallium chalcogenide cubane compounds are given in [\[link\]](#) and [\[link\]](#).



A plot of calculated vapor pressure (Torr) against temperature (K) for $[(\text{Me}_3\text{C})\text{GaS}]_4$. Adapted from E. G. Gillan, S. G. Bott, and A. R. Barron, *Chem. Mater.*, 1997, **9**, 3, 796.

The TGA approach to show reasonable agreement with previous measurements. For example, while the value calculated for $\text{Fe}(\text{acac})_3$ (2.78 Torr @ 113 °C) is slightly higher than that measured directly by the isoteniscopic method (0.53 Torr @ 113 °C); however, it should be noted that measurements using the sublimation bulb method obtained values much lower (8×10^{-3} Torr @ 113 °C). The TGA method offers a suitable alternative to conventional (direct) measurements of vapor pressure.

Bibliography

- P. W. Atkins, *Physical Chemistry*, 5th ed., W. H. Freeman, New York (1994).
- G. Beech and R. M. Lintonbon, *Thermochim. Acta*, 1971, **3**, 97.

- B. D. Fahlman and A. R. Barron, *Adv. Mater. Optics Electron.*, 2000, **10**, 223.
- E. G. Gillan, S. G. Bott, and A. R. Barron, *Chem. Mater.*, 1997, **9**, 3, 796.
- J. O. Hill and J. P. Murray, *Rev. Inorg. Chem.*, 1993, **13**, 125.
- J. P. Murray, K. J. Cavell and J. O. Hill, *Thermochim. Acta*, 1980, **36**, 97.
- M. A. V. Ribeiro da Silva and M. L. C. C. H. Ferrao, *J. Chem. Thermodyn.*, 1994, **26**, 315.
- R. Sabbah, D. Tabet, S. Belaadi, *Thermochim. Acta*, 1994, **247**, 193.
- L. A. Torres-Gomez, G. Barreiro-Rodriquez, and A. Galarza-Mondragon, *Thermochim. Acta*, 1988, **124**, 229.

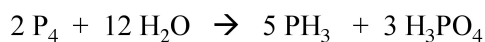
Phosphine and Arsine

Because of their use in metal organic chemical vapor deposition (MOCVD) of 13-15 (III-V) semiconductor compounds phosphine (PH₃) and arsine (AsH₃) are prepared on an industrial scale.

Synthesis

Phosphine (PH₃) is prepared by the reaction of elemental phosphorus (P₄) with water, [\[link\]](#). Ultra pure phosphine that is used by the electronics industry is prepared by the thermal disproportionation of phosphorous acid, [\[link\]](#).

Equation:

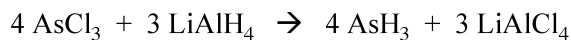


Equation:

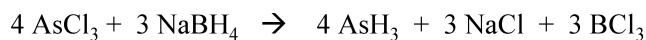


Arsine can be prepared by the reduction of the chloride, [\[link\]](#) or [\[link\]](#). The corresponding syntheses can also be used for stibine and bismuthine.

Equation:



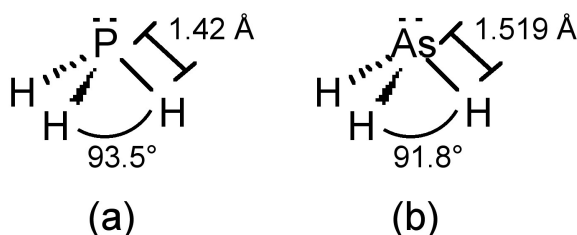
Equation:



The hydrolysis of calcium phosphide or arsenide can also generate the trihydrides.

Structure

The phosphorus in phosphine adopts sp^3 hybridization, and thus phosphine has an umbrella structure ([link](#)a) due to the stereochemically active lone pair. The barrier to inversion of the umbrella ($E_a = 155$ kJ/mol) is much higher than that in ammonia ($E_a = 24$ kJ/mol). Putting this difference in context, ammonia's inversion rate is 10^{11} while that of phosphine is 10^3 . As a consequence it is possible to isolate chiral organophosphines (PRR'R"). Arsine adopts the analogous structure ([link](#)b).



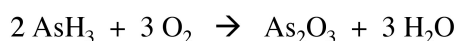
The structures of (a) phosphine and (b) arsine.

Reactions

Phosphine is only slightly soluble in water (31.2 mg/100 mL) but it is readily soluble in non-polar solvents. Phosphine acts as neither an acid nor a base in water; however, proton exchange proceeds via the phosphonium ion (PH_4^+) in acidic solutions and via PH_2^- at high pH, with equilibrium constants $K_b = 4 \times 10^{-28}$ and $K_a = 41.6 \times 10^{-29}$, respectively.

Arsine has similar solubility in water to that of phosphine (i.e., 70 mg/100 mL), and AsH_3 is generally considered non-basic, but it can be protonated by superacids to give isolable salts of AsH_4^+ . Arsine is readily oxidized in air, [link](#).

Equation:



Arsine will react violently in presence of strong oxidizing agents, such as potassium permanganate, sodium hypochlorite or nitric acid. Arsine decomposes to its constituent elements upon heating to 250 - 300 °C.

Gutzeit test

The Gutzeit test is the characteristic test for arsenic and involves the reaction of arsine with Ag^+ . Arsine is generated by reduction of aqueous arsenic compounds, typically arsenites, with Zn in the presence of H_2SO_4 . The evolved gaseous AsH_3 is then exposed to silver nitrate either as powder or as a solution. With solid AgNO_3 , AsH_3 reacts to produce yellow Ag_4AsNO_3 , while with a solution of AgNO_3 black Ag_3As is formed.

Hazards

Pure phosphine is odorless, but technical grade phosphine has a highly unpleasant odor like garlic or rotting fish, due to the presence of substituted phosphine and diphosphine (P_2H_4). The presence of P_2H_4 also causes spontaneous combustion in air. Phosphine is highly toxic; symptoms include pain in the chest, a sensation of coldness, vertigo, shortness of breath, and at higher concentrations lung damage, convulsions and death. The recommended limit (RL) is 0.3 ppm.

Arsine is a colorless odorless gas that is highly toxic by inhalation. Owing to oxidation by air it is possible to smell a slight, garlic-like scent when arsine is present at about 0.5 ppm. Arsine attacks hemoglobin in the red blood cells, causing them to be destroyed by the body. Further damage is caused to the kidney and liver. Exposure to arsine concentrations of 250 ppm is rapidly fatal: concentrations of 25 – 30 ppm are fatal for 30 min exposure, and concentrations of 10 ppm can be fatal at longer exposure times. Symptoms of poisoning appear after exposure to concentrations of 0.5 ppm and the recommended limit (RL) is as low as 0.05 ppm.

Bibliography

- R. Minkwitz, A. Kornath, W. Sawodny, and H. Härtner, *Z. Anorg. Allg. Chem.*, 1994, **620**, 753.

Mechanism of the Metal Organic Chemical Vapor Deposition of Gallium Arsenide

Introduction

Preparation of epitaxial thin films of III-V (13-15) compound semiconductors (notably GaAs) for applications in advanced electronic devices became a realistic technology through the development of metal organic chemical vapor deposition (MOCVD) processes and techniques. The processes mainly involves the thermal decomposition of metal alkyls and/or metal hydrides.

In 1968 Manasevit at the Rockwell Corporation was the first to publish on MOCVD for the epitaxial growth of GaAs. This followed his pioneering work in 1963 with the epitaxial growth of silicon on sapphire. The first publication used triethylgallium $[\text{Ga}(\text{CH}_2\text{CH}_3)_3]$ and arsine (AsH_3) in an open tube with hydrogen as the carrier gas. Manasevit actually coined the phrase MOCVD and since this seminal work there have been numerous attempts to improve and expand MOCVD for the fabrication of GaAs.

Several processes, partly in series, partly in parallel take place during the growth by CVD. They are presented schematically in [\[link\]](#). The relative importance of each of them depends on the chemical nature of the species involved and the design of the reactor used. The actual growth rate is determined by the slowest process in the series of events needed to come to deposition.

[missing_resource: GaAs Fig 1.jpg]

Schematic representation of the fundamental transport and reaction steps underlying MOCVD. Adapted from K. F. Jensen and W. Kern, in Thin Film Processes II, Eds. J. L. Vossen and W. Kern,

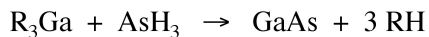
Academic Press, New York
(1991).

Conventionally, the metal organic chemical vapor deposition (MOCVD) growth of GaAs involves the pyrolysis of a vapor phase mixture of arsine and, most commonly, trimethylgallium [Ga(CH₃)₃, TMG] and triethylgallium [Ga(CH₂CH₃)₃, TEG]. Traditionally, growth is carried out in a cold-wall quartz reactor in flowing H₂ at atmospheric or low pressure. The substrate is heated to temperatures 400 - 800 °C, typically by RF heating of a graphite susceptor. Transport of the metal-organics to the growth zone is achieved by bubbling a carrier gas (e.g., H₂) through the liquid sources that are in held temperature-controlled bubblers.

Reaction mechanism

While the overall reaction (where R = CH₃ or CH₂CH₃) can be described by [\[link\]](#).

Equation:



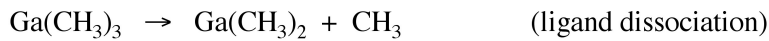
The nature of the reaction is much more complex. From early studies it was thought that free Ga atoms are formed by pyrolysis of TMG and As₄ molecules are formed by pyrolysis of AsH₃ and these species recombine on the substrate surface in an irreversible reaction to form GaAs.

Although a Lewis acid-base complex formed between TMG and AsH₃ is possible, it is now known that if there is any intermediate reaction between the TMG and AsH₃, the product is unstable. However, early work indicated that free GaAs molecules resulted from the decomposition of a TMG.AsH₃ intermediate and that the heated surface contributed to the reaction. It was subsequently suggested that the reaction occurs by separate pyrolysis of the reactants and a combination of individual Ga and As atoms at the surface or just above it. Finally, evidence has also been found for TMG pyrolysis

followed by diffusion through a boundary layer and for AsH₃ pyrolysis catalyzed by the GaAs surface.

There are several different kinds of potential reactions occurring in the CVD reaction chamber, namely, ligand dissociation, ligand association, , reductive elimination, oxidative addition, β-hydride elimination, etc. Some of them are listed in the following equations:

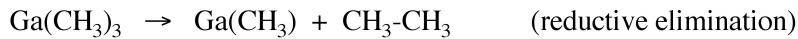
Equation:



Equation:



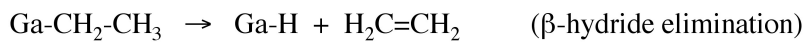
Equation:



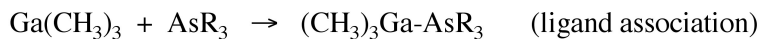
Equation:



Equation:



Equation:



Using ALE studies as insight for MOCVD

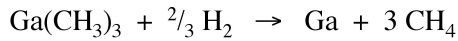
Given the stepwise and presumably simplified mechanism for atomic layer epitaxy (ALE) growth of GaAs, a number of mechanistic studies have been undertaken of ALE using TMG and AsH₃ to provide insight into the comparable MOCVD reactions. Nishizawa and Kurabayashi proposed that a CH₃-terminated GaAs surface inhibits further heterogeneous decomposition of TMG and self-limits the growth rate to one monolayer/cycle. While, X-ray photoelectron spectroscopy (XPS) studies showed that no carbon was observed on a GaAs surface reacted with TMG. Furthermore, the same self-limiting growth was seen in ALE using a metalorganic molecular beam epitaxy (MOMBE) with TMG and AsH₃. It was reported that a transient surface reconstruction is observable by reflection high-energy electron diffraction (RHEED) during the ALE of GaAs in MOMBE. It was suggested that this structure is caused by CH₃-termination and the self-limitation of the growth rate is attributed to this structure. However, measurement of the desorption of CH₃ by means of a combination of pulsed molecular beams and time-resolved mass spectrometry, indicates that CH₃ desorption is too fast to attribute the self-limitation to the CH₃-terminated surface. Subsequently, investigations of the pyrolysis of TMG on a (100)GaAs surface by the surface photo-absorption method (SPA) allowed for the direct observation of CH₃ desorption from a GaAs surface reacted with TMG. From the measured CH₃ desorption kinetics, it was shown that the CH₃-terminated surfaces causes the self-limitation of the growth rate in ALE because the excess TMG cannot adsorb.

All this research helped people to visualize the real reaction mechanism in the formation of GaAs by MOCVD methods, in which the decomposition, diffusion and surface reaction interact with each other and result in a much more complicated reaction mechanism.

Gas phase reaction: pyrolysis of TMG and AsH₃

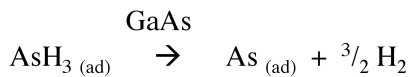
In the TMG/H₂ system, there is almost no reactions at a temperature below 450 °C, whereas the reaction of TMG with H₂ almost completely changed into CH₄ and Ga at a temperature above 600 °C, [\[link\]](#).

Equation:



As for the AsH_3 decomposition, without any deposition of Ga or GaAs in the reactor, the pyrolysis of AsH_3 proceeded barely at a temperature below 600 °C, however, it proceeded nearly completely at a temperature above 750 °C. In the AsH_3/H_2 system with the TMG introduced previously, the decomposition of AsH_3 was largely enhanced even at a temperature below 600 °C. The decomposition of AsH_3 seems to be affected sensitively by the deposited GaAs or Ga. This phenomenon may be concluded to be caused by the catalytic action by GaAs or Ga. The reaction at a temperature below 600 °C can be described as shown in [\[link\]](#), but at a temperature above 600 °C, pyrolysis of AsH_3 can occur even without GaAs or Ga, [\[link\]](#).

Equation:



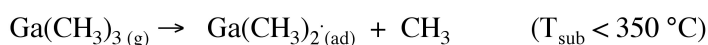
Equation:

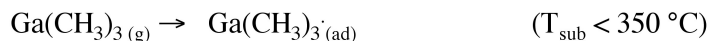


Adsorption and surface reactions

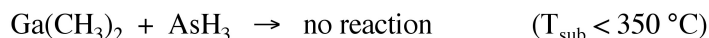
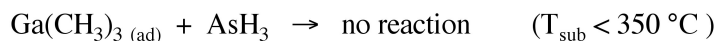
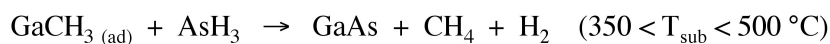
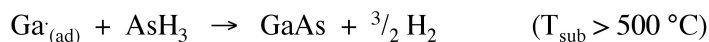
From the temperature dependent measurements of the desorption spectrum from a surface on which TMG was supplied, it was estimated that the surface-adsorbed species was Ga at the high temperature region of $T_{\text{sub}} > 500$ °C, GaCH_3 at the range of 350 °C $< T_{\text{sub}} < 500$ °C, and $\text{Ga}(\text{CH}_3)_2$ and $\text{Ga}(\text{CH}_3)_3$ at the range of $T_{\text{sub}} < 350$ °C. The reactions, where (ad) means the adsorbed state of the molecules, are:

Equation:



Equation:**Equation:****Equation:**

When AsH₃ is supplied, the reactions with these adsorbates are:

Equation:**Equation:****Equation:****Equation:**

It was observed that there is no growth in the range of $T_{\text{sub}} < 350 \text{ }^\circ\text{C}$, i.e., $\text{Ga}(\text{CH}_3)_2 (\text{ad})$ and $\text{Ga}(\text{CH}_3)_3 (\text{ad})$ do not react with AsH₃ in the TMG-AsH₃ system. Monomolecular layer growth is limited by the formation of GaCH₃ and its reaction with AsH₃.

Overall reaction pathway

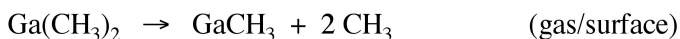
At lower temperature (350 - 500 °C), equivalently low energy, TMG decompose in the gas phase to $\text{Ga}(\text{CH}_3)_2$ and methyl radical, [\[link\]](#).

Equation:

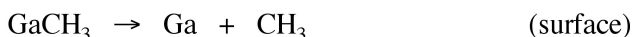


After the first ligand dissociation, there are two different pathways, in the first, the $\text{Ga}(\text{CH}_3)_2$ keeps decomposing into GaCH_3 and another methyl group when it is at the gas-substrate interface, [\[link\]](#), and then further decomposes into free gallium atoms on the substrate surface, [\[link\]](#). In the second reaction, the $\text{Ga}(\text{CH}_3)_2$ decomposes directly into Ga and $\text{CH}_3\text{-CH}_3$ by reductive elimination, [\[link\]](#).

Equation:



Equation:

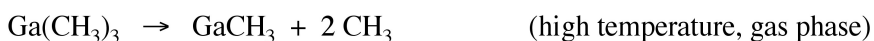


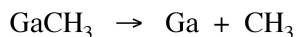
Equation:



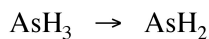
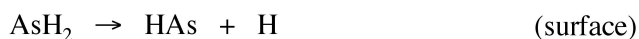
At high temperature (> 500 °C), the TMG decomposes into $\text{Ga}(\text{CH}_3)$ and two methyl groups instead of the step-wise decomposition at lower temperature, [\[link\]](#), and the $\text{Ga}(\text{CH}_3)$ further decomposes into free Ga atoms at the substrate surface, [\[link\]](#).

Equation:



Equation:

The decomposition of AsH_3 forms an “arsenic cloud” in the reaction chamber. The decomposition is also step-wise:

Equation:**Equation:****Equation:**

The methyl groups in the surface $\text{Ga}(\text{CH}_3)$ molecules are removed by the formation of methane with atomic hydrogen from the decomposition of AsH_3 , [\[link\]](#).

Equation:**Kinetics for other systems**

Investigations have been reported for the mechanism of the growth of GaAs using triethylgallium [$\text{Ga}(\text{CH}_2\text{CH}_3)_3$, TEG] and TMG with trimethylarsene [$\text{As}(\text{CH}_3)_3$, TMA], triethylarsene [$\text{As}(\text{CH}_2\text{CH}_3)_3$, TEA], *tert*-butylarsine $\{[(\text{CH}_3)_3\text{C}]\text{AsH}_2$, TBA}, and phenylarsine $[(\text{C}_6\text{H}_5)\text{AsH}_2]$. The experiments

were conducted in a MOCVD reactor equipped with a recording microbalance for in-situ growth rate measurements. For example, the kinetics of the growth of GaAs were investigated by measuring growth rate as a function of temperature using the microbalance reactor while holding the partial pressure of gallium precursor (e.g., TMG) and arsenic precursor [e.g., $\text{As}(\text{CH}_3)_3$] constant at 0.01 and 0.05 Torr, respectively. Three different flow rates were used to determine the influence of the gas residence time.

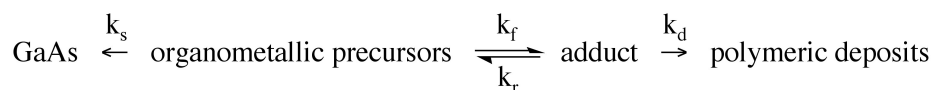
The growth rate of GaAs with TMG and $\text{As}(\text{CH}_2\text{CH}_3)_3$ is higher as compared with the growth from TMG and $\text{As}(\text{CH}_3)_3$ because of the lower thermal stability of $\text{As}(\text{CH}_2\text{CH}_3)_3$ than $\text{As}(\text{CH}_3)_3$. Both of the two growth rates showed a strong dependence on the residence time.

Similarly, the kinetic behaviors of the TMG/TBA and TEG/TBA system were investigated under the same conditions as the TMA and TEA studies. There are two distinct regions of growth. For TMG/TBA, the deposition rate is independent at low temperature and in the intermediate temperature (around 600 °C) the dependence of the growth rate on the total flow rate is significant. This means that the growth at the lower temperature is controlled by surface reactions. The TEG/TBA system showed a similar behavior except that the maximum growth rate occurs around 450 °C while it is around 750 °C for TMG/TBA system. Also, the growth of TMG/ $(\text{C}_6\text{H}_5)\text{AsH}_2$ was studied on the same conditions as for the $\text{Me}_3\text{Ga}/^t\text{BuAsH}_2$ system. It was reported that the difference in the growth rate at various flow rates was related to a combination of parasitic reactions and depletion effects from deposition. From the comparison of the data, it is deduced that the effect of parasitic reactions is slightly smaller for $(\text{C}_6\text{H}_5)\text{AsH}_2$ than for TBA.

Two possible mechanisms for the dependence of growth rate on flow rate were proposed. The first, mass-transfer limitation was thought to be unlikely because of the high diffusivity of the gallium precursors at 1 Torr (ca. 350 cm^2/s). The second, also the more likely explanation for the observed growth-rate dependence on flow rates is gas-phase depletion caused by the parasitic reactions. Since the growth efficiency is high (41% at 700 °C), the loss of precursor from the gas phase will directly affect the growth rate. It was evidenced by the differences in the growth rates between split

and combined feed streams. The growth rate is lower when the reagents are combined upstream of the reactor than when they are combined inside the reactor (split stream). It is suggested that the experimental observations can be explained by a model based on the reversible formation of an adduct and the decomposition of this adduct to useless polymeric material competing with the growth of GaAs. It can be written in the form shown in [\[link\]](#) where k_f and k_r are the forward and reverse rate constants for adduct formation, respectively, k_d is the rate constant for the irreversible decomposition of the adduct to polymer, and k_s is the surface reaction rate constant for the growth of GaAs. It is obvious that each step involves several elementary reactions, but there were insufficient data to provide any more detail.

Equation:



Bibliography

- T. H. Chiu, *Appl. Phys. Lett.*, 1989, **55**, 1244.
- H. Ishii, H. Ohno, K. Matsuzaki and H. Hasegawa, *J. Crys. Growth*, 1989, **95**, 132.
- K. F. Jensen and W. Kern, in *Thin Film Processes II*, Eds. J. L. Vossen and W. Kern, Academic Press, New York (1991).
- N. Kobayashi, Y. Yamauchi, and Y. Horikoshi, *J. Crys. Growth*, 1991, **115**, 353.
- K. Kodama, M. Ozeki, K. Mochizuki, and N. Ohtsuka, *Appl. Phys. Lett.*, 1989, **54**, 656.
- M. R. Leys and H. Veenvliet, *J. Crys. Growth*, 1981, **55**, 145.
- U. Memmert and M. L. Yu, *Appl. Phys. Lett.*, 1990, **56**, 1883.
- J. Nishizawa and T. Kurabayashi, *J. Crys. Growth*, 1988, **93**, 132.
- T. R. Omstead and K. F. Jensen, *Chem. Mater.*, 1990, **2**, 39.
- D. J. Schyer and M. A. Ring, *J. Electrochem. Soc.*, 1977, **124**, 569.
- Watanabe, T. Isu, M. Hata, T. Kamijoh, and Y. Katayama, *Japan. J. Appl. Phys.*, 1989, **28**, L1080.
- Y. Zhang, Th. Beuermann, and M. Stuke, *Appl. Phys. B*, 1989, **48**, 97.

- Y. Zhang, W. M. Cleaver, M. Stuke, and A. R. Barron, *Appl. Phys. A*, 1992, 55, 261.

Chemical Vapor Deposition of Silica Thin Films

General considerations

Before describing individual chemical vapor deposition (CVD) systems for the deposition of silica thin films, it is worth outlining general considerations to be taken into account with regard to the growth by CVD of any insulating film: the type of CVD method, deposition variables, and limitations of the precursor.

Deposition methods

In regard to the CVD of insulating films in general, and silica films in particular, three general reactors are presently used: atmospheric pressure CVD (APCVD), low and medium temperature low pressure CVD (LPCVD), and plasma-enhanced CVD (PECVD). LPCVD is often further divided into low and high temperatures.

APCVD systems allow for high throughput and even continuous operation, while LPCVD provides for superior conformal step coverage and better film homogeneity. PECVD has been traditionally used where low temperatures are required, however, film quality is often poor. As compared to PECVD, photo-assisted CVD has the additional advantage of highly selective deposition, although it has been little used in commercial systems. [\[link\]](#) summarizes the advantages and disadvantages of each type of CVD system commercially used for SiO₂ films.

	Atmospheric pressure CVD	Low temperature LPCVD	Medium temperature LPCVD	Plasma enhanced CVD
Temperature (°C)	300 - 500	300 - 500	500 - 900	100 - 350
Throughput	high	high	high	low
Step coverage	poor	poor	conformal	poor
Film properties	good	good	excellent	poor
Uses	passivation, insulation	passivation, insulation	insulation	passivation, insulation

Comparison of different deposition methods for SiO₂ thin films.

Deposition variables

The requirements of CVD films for electronic device applications have become increasingly more stringent as device sizes are continually reduced. Film thickness must be uniform across an entire wafer, i.e., better than $\pm 1\%$. The structure of the film and its composition must be controlled and reproducible, both on a single wafer, as well as between wafer samples. It is also desirable that the process is safe, inexpensive, and easily automated.

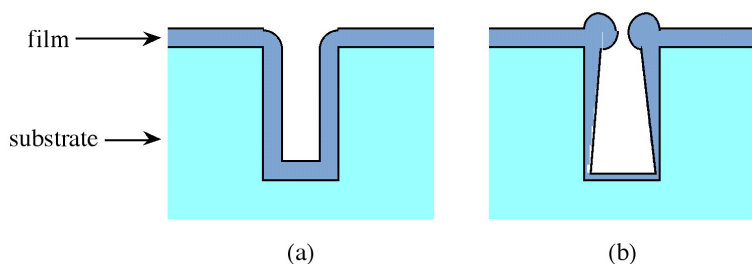
A number of variables determine the quality and rate of film growth for any material. In general, the deposition rate increases with increased temperature and follows the Arrhenius equation, [\[link\]](#), where R is the deposition rate, E_a is the activation energy, T is the temperature (K), A is the frequency factor, and k is Boltzmann's constant (1.381×10^{-23} J/K).

Equation:

$$R = A \exp(-E_a/kT)$$

At the high temperatures the rate of deposition becomes mass transport limited. Meaning, the rate of surface reaction is faster than the rate at which precursors are transported to the surface. In multiple source systems, the film growth rate is dependent on the vapor phase concentration (or partial pressure) of each of the reactants, but in certain cases the ratio of reactants is also important, e.g., the SiH₄/O₂ growth of SiO₂. Surface catalyzed reactions can also alter the deposition rate. Such as the non-linear dependence of the deposition rate of SiO₂ on the partial pressure of Si(OEt)₄. Gas depletion may also be significant requiring either a thermal ramp in the chamber and/or special reactor designs. The necessary incorporation of dopants usually lowers deposition rates, due to competitive surface binding.

For the applications of insulating materials as isolation layers, an important consideration is step coverage: whether a coating is uniform with respect to the surface. [\[link\]](#)a shows a schematic of a completely uniform or conformal step coverage of a trench (such as occurs between isolated devices) where the film thickness along the walls is the same as the film thickness at the bottom of the step. Uniform step coverage results when reactants or reactive intermediates are able to migrate rapidly along the surface before reacting. When the reactants adsorb and react without significant surface migration, deposition is dependent on the mean free path of the gas. [\[link\]](#)b shows an example of minimal surface migration and a short mean free path. For SiO₂ film growth LPCVD has highly uniform coverage ([\[link\]](#)a) and PECVD poor step coverage ([\[link\]](#)b).



Step coverage of deposited films with (a) uniform coverage resulting from rapid surface migration and (b) nonconformal step coverage due to no surface migration.

Precursor considerations

The general requirements for any CVD precursor have been adequately reviewed elsewhere, and will not be covered here. However, many of the gases and organometallics used to deposit dielectric films are hazardous. The safety problems are more severe for LPCVD because the process often uses no diluent gas such as argon or nitrogen. [\[link\]](#) lists the boiling point and hazards of common inorganic and organometallic precursor sources for CVD of SiO_2 and doped silica. Many of the precursors react with air to form solid products, thus leaks can cause particles to form in the chamber and gas lines.

Gas	Formula	Bpt (°C)	Hazard
ammonia	NH_3	-33.35	toxic, corrosive
argon	Ar	-185.7	inert
arsine	AsH_3	-55	toxic
diborane	B_2H_6	-92.5	toxic, flammable
dichlorosilane	SiCl_2H_2	8.3	toxic, flammable
hydrogen	H_2	-252.8	flammable
nitrogen	N_2	-209.86	inert

nitrous oxide	N ₂ O	-88.5	oxidizer
oxygen	O ₂	-182.962	oxidizer
phosphine	PH ₃	-87.7	toxic, P ₂ H ₄ impurities, flammable
silane	SiH ₄	-111.8	flammable, toxic

Physical and hazard properties of common gaseous sources for CVD of dielectric materials.

In principle, the deposition of a SiO₂, or silica, thin film by CVD requires two chemical sources: the element (or elements) in question, and an oxygen source. While dioxygen (O₂) is suitable for many applications, its reactions may be too fast or too slow for optimum film growth, requiring that alternative oxygen sources be used, e.g., nitrous oxide (N₂O) and ozone (O₃). A common non-oxidizing oxygen source is water. A more advantageous approach is to incorporate oxygen into the ligand environment of the precursor, and endeavor to preserve such an interaction intact from the source molecule into the ultimate film; such a source is often termed a "single-source" precursor.

CVD silica (SiO₂)

The processing sequence for silicon dioxide (SiO₂) used depends on its specific use. CVD processes for SiO₂ films can be characterized by either the chemical reaction type, the growth pressure, or the deposition temperature. The choice of route is often dictated by requirements of the thermal stability of the substrate or the conformality. [\[link\]](#) summarizes selected properties of SiO₂ grown by various CVD methods, in comparison to that of thermally grown silica. In general, silica grown at high temperatures resemble thermally grown "native" SiO₂. However, the use of aluminum metallization requires low temperature deposition of silica.

Deposition	Plasma	SiH ₄ + O ₂	Si(OEt) ₄	SiCl ₂ H ₂ + N ₂ O	Thermal
Temperature (°C)	200	450	700	900	1000
Composition	SiO _{1.9} (H)	SiO ₂ (H)	SiO ₂	SiO ₂ (Cl)	SiO ₂
Step coverage	non-conformal	non-conformal	conformal	conformal	conformal
Thermal stability	loses H	densifies	stable	loses Cl	stable

Refractive index	1.47	1.44	1.46	1.46	1.46
Dielectric constant	4.9	4.3	4.0	4.0	3.9

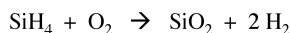
Comparison of physical properties of SiO₂ grown by commercial CVD methods.

CVD from hydrides

The most widely used method for SiO₂ thin film CVD is the oxidation of silane (SiH₄), first developed in 1967 for APCVD. Nonetheless, LPCVD systems have since become increasingly employed, and exceptionally high growth rates (30,000 Å/min) have been obtained by the use of rapid thermal CVD.

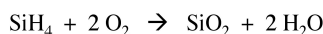
The chemical reaction for SiO₂ deposition from SiH₄ is:

Equation:



At high oxygen partial pressures an alternative reaction occurs, resulting in the formation of water.

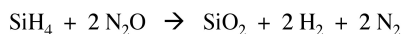
Equation:



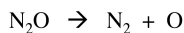
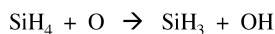
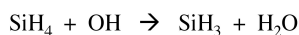
While these reactions appear simple, the detailed mechanism involves a complex branching-chain sequence of reactions. The apparent activation energy is low (< 41 kJ/mol) as a consequence of its heterogeneous nature, and involves both surface adsorption and surface catalysis.

Nitrous oxide (N₂O) can be used as an alternative oxygen source to O₂, according to the overall reaction, [\[link\]](#).

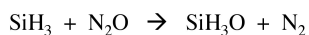
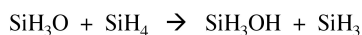
Equation:



A simple kinetic scheme has been developed to explain many of the observed aspects of SiH₄-N₂O growth. It was suggested that the reaction is initiated by decomposition of N₂O, [\[link\]](#), generating an oxygen radical which can abstract hydrogen from silane forming a hydroxyl radical, [\[link\]](#), that can react further with silane, [\[link\]](#).

Equation:**Equation:****Equation:**

Evidence for the reaction of the OH radical to form water is the formation of a small quantity of water observed during the oxidation of SiH₄. Silyl radicals are oxidized by N₂O to form siloxy radicals, [\[link\]](#), which provide a suitable propagation step, [\[link\]](#).

Equation:**Equation:**

It has been proposed that the silanol (SiH₃OH) is the penultimate film precursor.

The SiH₄-O₂ and SiH₄-N₂O routes to SiO₂ thin films are perhaps the most widely studied photochemical CVD system of all dielectrics. Photo-CVD of SiO₂ provides a suitable route to deposition at low substrate temperatures, thereby avoiding potential thermal effects of wafer warpage and deleterious dopant redistribution. In addition, unlike other low temperature methods such as APCVD and PECVD, photo-CVD often provides good purity of films.

A summary of common silane CVD systems is given in [\[link\]](#).

Oxygen source	Carrier gas (diluent)	CVD method	Deposition temp. (°C)	Growth rate (Å/min)
O ₂	N ₂	APCVD	350 - 475	100 - 14,000
O ₂	Ar	LPCVD	100 - 550	100 - 30,000

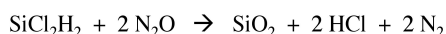
O ₂	Ar/N ₂	LPCVD	25 - 500	10 - 450
O ₂	Ar	PECVD	25 - 200	200 - 900
N ₂ O	N ₂	APCVD	490 - 690	200 - 1,200
N ₂ O	N ₂	LPCVD	700 - 860	<i>ca.</i> 50
N ₂ O	N ₂	LPCVD	25 - 350	7 - 180
N ₂ O	Ar	PECVD	100 - 200	80 - 800

Precursors and deposition conditions for SiO₂ CVD using silane (SiH₄).

CVD from halides

The most widely used process of the high temperature growth of SiO₂ by LPCVD involves the N₂O oxidation of dichlorosilane, SiCl₂H₂, [\[link\]](#).

Equation:



Deposition at 900 - 915 °C allows for growth of SiO₂ films at *ca.* 120 Å/min; however, these films are contaminated with Cl. Addition of small amounts of O₂ is necessary to remove the chlorine.

While PECVD has been employed utility halide precursors, the ability of small quantities of fluorine to improve the electrical properties of SiO₂ has prompted investigation of the use of SiF₄ as a suitable source.

CVD from tetraethoxysilane (TEOS)

The first CVD process to be introduced into semiconductor technology in 1961 was that involving the pyrolysis of tetraethoxysilane, Si(OEt)₄ (commonly called TEOS from tetraethylorthosilicate). Deposition occurs at an optimum temperature around 750 °C. However, under LPCVD conditions, the growth temperature can be significantly lowered (> 600 °C). The high temperature growth of SiO₂ from TEOS involves no external oxygen source. Dissociative adsorption studies indicate that decomposition of the TEOS-derived surface bound di- and tri-ethoxysiloxanes is the direct source of the ethylene.

PECVD significantly lowers deposition temperatures using TEOS, but requires the addition of O₂ to remove carbon contamination, via the formation of gaseous CO and CO₂, which are

subsequently not incorporated within the film. Although deposition as low as 100 °C may be obtained, the film resistivity increases by three orders of magnitude by depositing at 200 °C; being $10^{16} \Omega\cdot\text{cm}$, with a breakdown strength of $7 \times 10^6 \text{ V/cm}$.

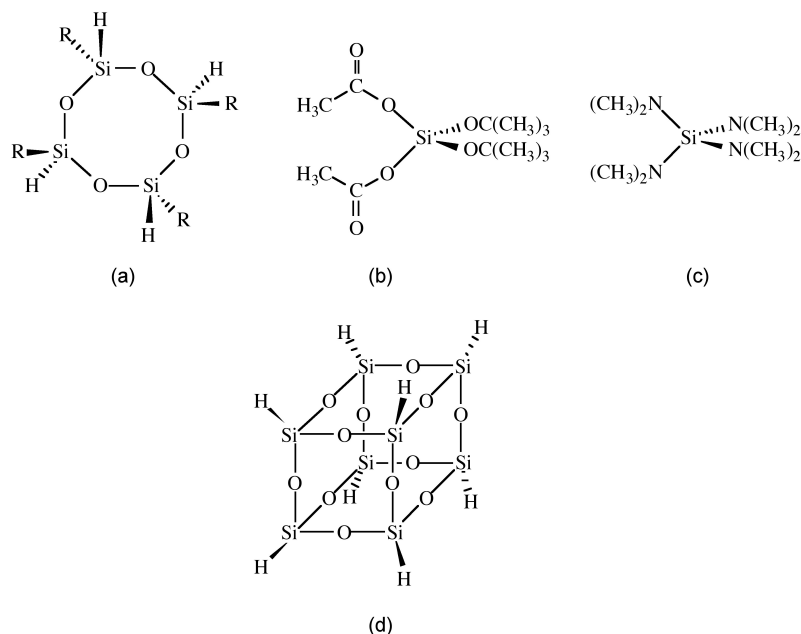
Addition of O_2 for APCVD growth does not decrease the deposition temperature, however, if ozone (O_3) is used as the oxidation source, deposition temperatures as low as 300 °C may be obtained for uniform crack-free films. It has been postulated that the ozone traps the TEOS molecule on the surface as it reacts with the ethoxy substituent, providing a lower energy pathway (TEOS- O_3 @ 55 kJ/mol versus TEOS- O_2 @ 230 kJ/mol and TEOS only @ 190 kJ/mol).

There are significant advantages of the TEOS/ O_3 system, for example the superior step coverage it provides. Furthermore, films have low stress and low particle contamination. On this basis the TEOS/ O_3 system has become widely used for silica, as well as silicate glasses.

CVD from other organosilicon precursors

A wide range of alternative silicon sources has been investigated, especially with regard to either lower temperature deposition and/or precursors with greater ambient stability.

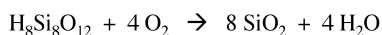
Diethylsilane (Et_2SiH_2), 1,4-disilabutane (DBS, $\text{H}_3\text{SiCH}_2\text{CH}_2\text{SiH}_3$), 2,4,6,8-tetramethylcyclotetrasiloxane (TMCTS, [\[link\]](#)a, where $\text{R} = \text{CH}_3$), and 2,4,6,8-tetraethylcyclotetrasiloxane (TECTS, [\[link\]](#)a, where $\text{R} = \text{C}_2\text{H}_5$), have been used in conjunction with O_2 over deposition temperatures of 100 - 600 °C, depending on the precursor. Diacetoxymethyl-*tert*-butyl silane (DADBS, [\[link\]](#)b) has been used without additional oxidation sources. High quality silicon oxide has been grown at 300 °C by APCVD using the amido precursor, $\text{Si}(\text{NMe}_2)_4$ ([\[link\]](#)c).



Alternative organometallic silicon sources that have been investigated for the growth of silica thin films.

An interesting concept has been to preform the -Si-O-Si- framework in the precursor. In this regard, the novel precursor T₈-hydridospherosiloxane (H₈Si₈O₁₂, [\[link\]](#)d) gives smooth amorphous stoichiometric SiO₂ at 450 - 525 °C by LPCVD. The decomposition mechanism in the presence of added oxygen involves the loss of water, [\[link\]](#). IR studies indicate that the Si-O-Si bonds are preserved during deposition. While films are of high quality, the present synthesis of H₈Si₈O₁₂ is of low yield (*ca.* 21%), making it currently impractical for large scale processing.

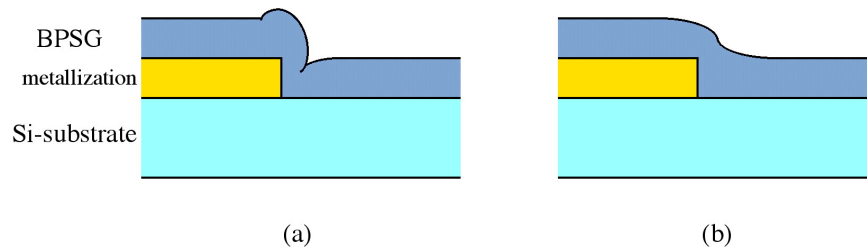
Equation:



CVD silicate glasses

Borosilicate glasses (BSG), phosphosilicate glasses (PSG) and borophosphosilicate glasses (BPSG) are frequently used as insulating layers separating conducting layers. These glasses have lower intrinsic stress, lower melting temperatures and better dielectric properties than SiO₂ itself. PSG and BPSG have the added property of gettering and immobilizing dopants. Particularly important is the gettering of sodium ions, which are a source of interface traps. The low temperature molten properties of BSG, PSG, and BPSG glasses allow for the smoothing of the device topography by viscous thermal fusion to convert abrupt steps to more gradually tapered steps ([\[link\]](#)a) as well as planarization of complex topologies ([\[link\]](#)b), enabling deposition of continuous metal layers. This process is commonly called P-glass flow. The boron

and phosphorous contents of the silicate glasses vary, depending on the application, typically being from 2 to 8 weight per cent.



Schematic cross section of BPSG as deposited (a) and after annealing (b), showing the flow causing a decrease in the angle of the BPSG going over the step.

The advantage of BPSG over PSG is that flow occurs over the temperature range of 750 - 950 °C, depending on the relative P and B content (as opposed to 950 - 1110 °C for PSG). Lowering of the flow temperature is required to minimize dopant migration in VLSI devices. Conversely, the disadvantages of BPSG versus PSG include the formation of bubbles of volatile phosphorous oxides and crystallites of boron-rich phases. If, however, the dopant concentration is controlled, these effects can be minimized.

Arsenosilicates (AsSG) were employed originally in silicon device technology as an arsenic dopant source for planar substrates prior to the advent of large scale ion implantation which has largely removed the need for AsSG in doping applications. However, with ULSI silicon circuit fabrication, the requirement for doping of deep trenches (inaccessible to ion implantation) has witnessed the re-emergence of interest in AsSG films.

The CVD growth of silicate glasses follows that of SiO₂, with SiH₄ and TEOS being the most commonly employed silicon precursors. A summary of common CVD precursor systems for silicate glasses is given in [\[link\]](#).

Precursors	CVD method	Deposition temp. (°C)	Applications
SiH ₄ /B ₂ H ₆	APCVD	300 - 450	good step coverage
SiH ₄ /B ₂ H ₆	LPCVD	350 - 400	-

SiH_4/PH_3	APCVD	300 - 450	-
SiH_4/PH_3	LPCVD	350 - 400	flow glass
$\text{SiH}_4/\text{B}_2\text{H}_6/\text{PH}_3$	APCVD	300 - 450	-
$\text{SiH}_4/\text{B}_2\text{H}_6/\text{PH}_3$	LPCVD	350 - 400	-
$\text{SiH}_4/\text{AsH}_3$	APCVD	500 - 700	-
$\text{TEOS}/\text{B}(\text{OMe})_3$	APCVD	650 - 730	diffusion source
$\text{TEOS}/\text{B}(\text{OMe})_3$	LPCVD	500 - 750	trench filling
$\text{TEOS}/\text{B}(\text{OEt})_3$	APCVD	475 - 800	diffusion source
$\text{TEOS}/\text{B}(\text{OEt})_3$	LPCVD	500 - 750	diffusion source
TEOS/PH_3	LPCVD	650	flow glass
$\text{TEOS}/\text{O}=\text{P}(\text{OMe})_3$	APCVD	300 - 800	flow glass
$\text{TEOS}/\text{P}(\text{OMe})_3$	LPCVD	500 - 750	diffusion source
$\text{TEOS}/\text{O}=\text{P}(\text{OMe})_3$	LPCVD	500 - 800	flow glass
$\text{TEOS}/\text{B}(\text{OMe})_3/\text{PH}_3$	LPCVD	620 - 800	trench filling
$\text{TEOS}/\text{B}(\text{OMe})_3/\text{P}(\text{OMe})_3$	LPCVD	675 - 750	flow glass
$\text{TEOS}/\text{B}(\text{OMe})_3/\text{O}=\text{P}(\text{OMe})_3$	LPCVD	680	flow glass
$\text{TEOS}/\text{AsCl}_3$	APCVD	500 - 700	diffusion source
$\text{TEOS}/\text{As}(\text{OEt})_3$	LPCVD	700 - 730	trench doping
$\text{TEOS}/\text{O}=\text{As}(\text{OEt})_3$	LPCVD	700 - 730	trench doping

Precursors and deposition conditions for CVD of borosilicate glass (BSG), phososilicate glass (PSG), borophosphosilicate glass (BPSG) and arsenosilicates (AsSG) thin films.

CVD from hydrides

Films of BSG, PSG, and BPSG may all be grown from SiH_4 , O_2 and B_2H_6 and/or PH_3 , at 300 - 650 °C. For APCVD, the reactants are diluted with an inert gas such as nitrogen, and the O_2 /hydride molar ratio is carefully controlled to maximize growth rate and dopant concentration (values of 1 to 100 are used depending on the application). Ordinarily, the dopant concentration for both BSG and PSG decreases with increased temperature. However, some reports indicate an increase in boron content with increased temperature. Film growth of BPSG was found to occur in two temperature regions. Deposition at low temperature (270 - 360 °C) occurred via a surface reaction rate limiting growth ($E_a = 39$ kcal/mol), while at higher temperature (350 - 450 °C), a mass-transport rate limited reaction region is observed ($E_a = 7.6$ kcal/mol).

LPCVD of BSG and PSG is conducted at 450 - 550 °C with an O_2 :hydride ratio of 1:1.5. Conversely, an O_2 :hydride ratio of 1.5:1 provides the optimum growth conditions for BPSG over the same temperature range. The phosphorous in PSG films was found to exist as a mixture of P_2O_5 and P_2O_3 , however, the latter can be minimized under the correct deposition conditions. Some difficulties have been reported for the use of B_2H_6 due to its thermal instability. Substitution of B_2H_6 with BCl_3 obviates this problem, although the resulting films are invariably contaminated with 1 weight per cent chloride.

Arsenosilicate glass (AsSG) thin films are generally grown by APCVD using arsine (AsH_3); the use of which is being limited due to its high toxicity. However, arsine inhibits the gas phase reactions between SiH_4 and O_2 , such that film grown from $\text{SiH}_4/\text{AsH}_3/\text{O}_2$ show improved step coverage at high deposition rates.

CVD from metal organic precursors

As with SiO_2 deposition, see above, there has been a trend towards the replacement of SiH_4 with TEOS on account of its ability to produce highly conformal coatings. This is particularly attractive with respect to trench filling. Furthermore, films of doped SiO_2 glasses have been obtained using both APCVD and LPCVD (typically below 3 Torr), with a wide variety of dopant elements including: boron, phosphorous, and arsenic, including antimony, tin, and zinc.

Boron-containing glasses are generally grown using either trimethylborate, $\text{B}(\text{OMe})_3$, or triethylborate, $\text{B}(\text{OEt})_3$, although the multi-element source, *tris*(trimethylsilyl)borate, $\text{B}(\text{OSiMe}_3)_3$, has been employed for both silicon and boron in BPSG thin film growth. Similarly, whereas PH_3 may be used as the phosphorous source, trimethylphosphite, $\text{P}(\text{OMe})_3$, and trimethylphosphate, $\text{O}=\text{P}(\text{OMe})_3$, are preferred. Likewise, triethoxyarsine, $\text{As}(\text{OEt})_3$, and triethylarsenate, $\text{O}=\text{As}(\text{OEt})_3$, have been employed for AsSG growth.

The co-reaction of TEOS with organoboron and organophosphorous compounds allows for deposition at lower temperatures (500 - 650 °C) than for hydride growth of comparable rates. However, LPCVD, using an all organometallic approach, requires $\text{P}(\text{OMe})_3$ because the low reactivity of $\text{O}=\text{P}(\text{OMe})_3$ prevents significant phosphorus incorporation. Although premature decomposition of $\text{P}(\text{OMe})_3$ occurs at 600 °C (leading to non-uniform growth), deposition at 550 °C results in high film uniformity at reasonable deposition rates.

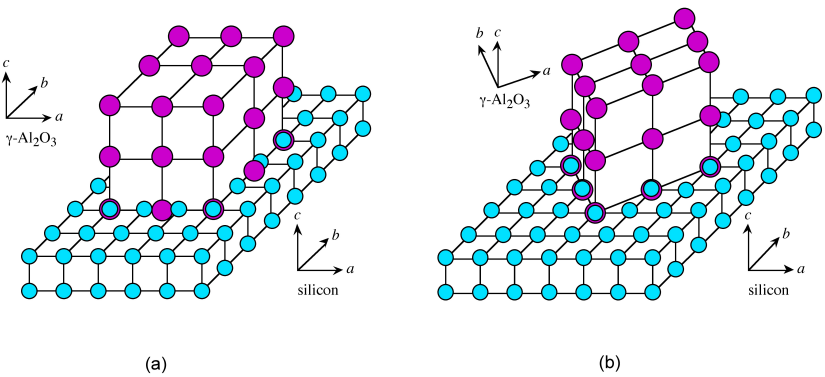
Bibliography

- W. Kern and V. S. Ban, in *Thin Film Processes*, Eds. J. L. Vossen, W. Kern, Academic Press, New York (1978).
- M. L. Hammod, *Solid State Technol.*, 1980, **23**, 104.
- A. R. Barron and W. S. Rees, Jr., *Adv. Mater. Optics Electron.*, 1993, **2**, 271.
- N. Goldsmith and W. Kern, *RCA Rev.*, 1967, **28**, 153.
- C. Pavelescu, J. P. McVittie, C. Chang, K. C. Saraswat, and J. Y. Leong, *Thin Solid Films*, 1992, **217**, 68.
- J. D. Chapple-Sokol, C. J. Giunta, and R. G. Gordon, *J. Electrochem. Soc.*, 1987, **136**, 2993.
- P. González, D. Fernández, J. Pou, E. García, J. Serra, B. León, and M. Pérez-Amor, *Thin Solid Films*, 1992, **218**, 170.
- E. L. Jordan, *J. Electrochem. Soc.*, 1961, **108**, 478.
- K. Fujino, Y. Nishimoto, N. Tokumasu, and K. Maeda, *J. Electrochem. Soc.*, 1990, **137**, 2883.
- R. A. Levy and K. Nassau, *J. Electrochem. Soc.*, 1986, **133**, 1417.
- L. K. White, J. M. Shaw, W. A. Kurylo, and N. Miskowski, *J. Electrochem. Soc.*, 1990, **137**, 1501.

Chemical Vapor Deposition of Alumina

Alumina

Alumina, Al_2O_3 , exists as multiple crystalline forms, however, the two most important are the α and γ forms. $\alpha\text{-Al}_2\text{O}_3$ (corundum) is stable at high temperatures and its structure consists of a hexagonal close-packed array of oxide (O^{2-}) ions with the Al^{3+} ions occupying octahedral interstices. In contrast, $\gamma\text{-Al}_2\text{O}_3$ has a defect spinel structure, readily takes up water and dissolves in acid. Despite the potential disadvantages of $\gamma\text{-Al}_2\text{O}_3$ there is a preference for its deposition on silicon substrates because of the two different lattice-matching relationships of $\gamma\text{-Al}_2\text{O}_3$ (100) on $\text{Si}(100)$. These are shown as schematic diagrams in [\[link\]](#). A summary of CVD precursor systems for Al_2O_3 is given in [\[link\]](#).



Schematic diagram of the crystallographic relations of $\gamma\text{-Al}_2\text{O}_3$ on $\text{Si}(100)$: (a) $\gamma\text{-Al}_2\text{O}_3$ (100)|| $\text{Si}(100)$, and (b) $\gamma\text{-Al}_2\text{O}_3$ (100)|| $\text{Si}(110)$. Adapted from A. R. Barron, *CVD of Non-Metals*, W. S. Rees, Jr., Ed. VCH, New York (1996).

Aluminum precursor	Oxygen source	Carrier gas	CVD method	Deposition temp. ($^{\circ}\text{C}$)	Comments
AlCl_3	CO_2/H_2	H_2 or N_2	APCVD	700 - 900	amorphous (700),

					crystalline (850 - 900)
AlMe_3	O_2	N_2 or He	APCVD	350 - 380	dep. rate highly dependent on gas- phase conc. Al and O_2
AlMe_3	O_2	N_2	LPCVD	375	plasma- enhanced, 10 W
AlMe_3	N_2O	N_2 or He	APCVD	100 - 660	lower quality than with O_2
AlMe_3	N_2O	N_2	LPCVD	950 - 1050	good passivation properties of Si MOS devices
AlMe_3	N_2O	He	PECVD	120 - 300	plasma- enhanced,
$\text{Al}(\text{O}^i\text{Pr})_3$	O_2	N_2	APCVD	420 - 600	
$\text{Al}(\text{O}^i\text{Pr})_3$	O_2	N_2	LPCVD	250 - 450	
$\text{Al}(\text{O}^i\text{Pr})_3$	N_2O	Ar	LPCVD	200 - 750	epitaxial on Si
$\text{Al}(\text{acac})_3$		N_2	APCVD	420 - 450	high C content
$\text{Al}(\text{acac})_3$	air	N_2	APCVD	250 - 600	significant C content
$\text{Al}(\text{acac})_3$	O_2 and H_2O	Ar	LPCVD	230 - 550	growth rate indep. of

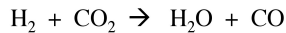
					H ₂ O but film quality dep. on H ₂ O
--	--	--	--	--	--

Precursors and deposition conditions for Al₂O₃ CVD.

CVD from halides

The initial use of CO₂/H₂ as a hydrolysis source for the CVD of SiO₂ from SiCl₄, led to the analogous deposition of Al₂O₃ from AlCl₃, i.e.,

Equation:



Deposition in the temperature range 700 - 900 °C was found to yield films with optimum dielectric properties, but films deposited below 700 °C contained significant chloride impurities. It has been determined that H₂O vapor, formed from H₂ and CO₂, acts as the oxygen donor, and not the CO₂. The crystal form of the CVD-grown alumina films was found to depend on the deposition temperature; films grown below 900 °C were γ-Al₂O₃, while those grown at 1200 °C were α-Al₂O₃, in accord with the known phase diagram for this material.

CVD from trimethylaluminum (TMA)

Although trimethylaluminum, AlMe₃ (TMA), reacts rapidly with water to yield Al₂O₃, the reaction is highly exothermic (-1243 kJ/mol) and thus difficult to control. The oxygen gettering properties of aluminum metal, however, can be employed in the controlled MOCVD growth of Al₂O₃. The common deposition conditions employed for CVD of Al₂O₃ from AlMe₃ are similar to those used for aluminum-metal CVD, but with the addition of an oxygen source, either O₂ or N₂O.

Films grown by APCVD using N₂O are of inferior quality to those employing O₂, due to their exhibiting some optical absorption in the visible wavelength region. The growth of high quality films using either oxygen source is highly dependent on the gas phase concentrations of aluminum and “oxygen”. Further improvements in film quality are observed with the use of a temperature gradient in the chambers deposition zone.

Attempts to lower the deposition temperature employing PECVD have been generally successful. However, a detailed spectroscopic study showed that the use of N_2O as the oxygen source resulted in significant carbon and hydrogen incorporation at low temperatures (120 - 300 °C). The carbon and hydrogen contamination are lowered at high deposition temperature, and completely removed by a post-deposition treatment under O_2 . It was proposed that the carbon incorporated in the films is in the chemical form of Al-CH_3 or Al-C(O)OH , while hydrogen exists as Al-OH moieties within the film.

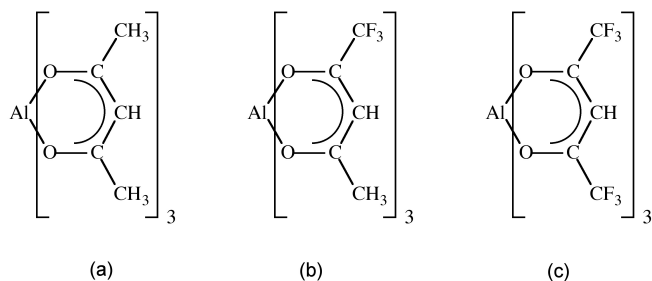
Photo-assisted CVD of Al_2O_3 from AlMe_3 has been reported to provide very high growth rates (2000 Å/min) and give films with electrical properties comparable to films deposited using thermal or plasma techniques. Irradiation with a 248 nm (KrF) laser source allowed for uniform deposition across a 3" wafer. However, use of 193 nm (ArF) irradiation required dilution of the AlMe_3 concentration to avoid non-uniform film growth.

CVD from alkoxides and β -diketonates

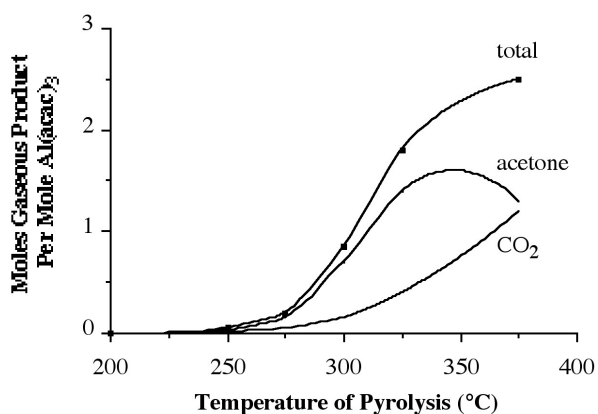
The pyrophoric nature of AlMe_3 urged investigations into alternative precursors, in particular those which already contain oxygen. Alternative precursors might also provide possible routes to eliminate carbon contamination. Given the successful use of TEOS in SiO_2 thin film growth, an analogous alkoxide precursor approach is logical. The first report of Al_2O_3 films grown by CVD used an aluminum alkoxide precursors.

Aluminum tris-*iso*-propoxide, $\text{Al}(\text{O}^i\text{Pr})_3$, is a commercially available inexpensive alkoxide precursor compound. Deposition may be carried-out by either APCVD or LPCVD, using oxygen as an additional oxidation source to ensure low carbon contamination. It is adventitious to use LPCVD (10 Torr) growth to inhibit gas phase homogeneous reactions, causing formation of a powdery deposit. The use of lower chamber pressures (3 Torr) and N_2O as the oxide source provided sufficient improvement in film quality to allow for device fabrication.

The deposition of Al_2O_3 films from the pyrolysis of aluminum acetylacetonate, $\text{Al}(\text{acac})_3$ ([link](#)), has been widely investigated using both APCVD and LPCVD. The perceived advantage of $\text{Al}(\text{acac})_3$ over other aluminum precursors includes lowered-toxicity, good stability at room temperature, easy handling, high volatility at elevated temperatures, and low cost. However, the quality of films was originally poor; carbon being the main contaminant resulting from the thermolysis and incorporation of acetone and carbon dioxide formed upon thermal decomposition ([link](#)).



Aluminum β -diketonate precursors.



Gaseous decomposition products from the pyrolysis of $\text{Al}(\text{acac})_3$ as a function of pyrolysis temperature (Data from J. Von Hoene, R. G. Charles, and W. M. Hickam, *J. Phys. Chem.*, 1958, **62**, 1098).

Incomplete oxidation of the film may be readily solved by the addition of water vapor to the carrier gas stream; pure carbon-free films being grown at temperatures as low as 230 °C. In fact, water vapor plays an important role in the film growth kinetics, film purity, and the surface morphology of the grown films. While the growth rate is unaffected by the addition of water vapor, its influence on the surface morphology is significant. Films grown without water vapor on the Al_2O_3 surface is rough with particulates. In contrast, films grown with water vapor are mirror smooth.

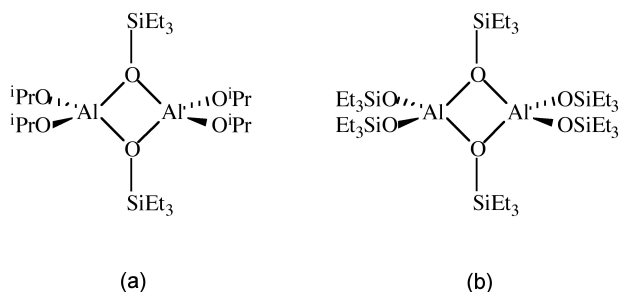
A systematic study of the kinetics of vaporization of $\text{Al}(\text{acac})_3$ along with fluorinated aluminum β -diketonate complexes, $\text{Al}(\text{tfac})_3$ ([link](#)b) and $\text{Al}(\text{hfac})_3$ ([link](#)c), has been reported, and the saturation vapor pressures determined at 75 - 175 °C.

Aluminum silicates

The high dielectric constant, chemical stability and refractory character of aluminosilicates, $(\text{Al}_2\text{O}_3)_x(\text{SiO}_2)_y$, makes them useful as packaging materials in IC chip manufacture. Mullite ($3\text{Al}_2\text{O}_3 \cdot 2\text{SiO}_2$) prepared by sol-gel techniques, is often used as an encapsulant for active devices and thin-film components. Amorphous alumina-silica films have also been proposed as insulators in multilevel interconnections, since they do not suffer the temperature instability of alumina films retain the desirable insulating characteristics. Under certain conditions of growth and fabrication, silica may crystallize, thereby allowing diffusion of oxygen and impurities along grain boundaries to the silicon substrate underneath. Such unwanted reactions are catastrophic to the electronic properties of the device. The retention of amorphous structure over a larger temperature range of silicon rich alumina-silica films offers a possible solution to this deleterious diffusion.

Thin films of mixed metal oxides are usually obtained from a mixture of two different kinds of alkoxide precursors. However, this method suffers from problems with stoichiometry control since extensive efforts must be made to control the vapor phase concentration of two precursors with often dissimilar vapor pressures. Also of import here is the near impossible task of matching rates of hydrolysis/oxidation to give "pure", non-phase segregated films, i.e., those having a homogeneous composition and structure. In an effort to solve these problems, research effort has been aimed at single-source precursors, i.e., those containing both aluminum and silicon.

The first study of single-source precursors for $(\text{Al}_2\text{O}_3)_x(\text{SiO}_2)_y$ films employed the mono-siloxide complex $\text{Al}(\text{O}^i\text{Pr})_2(\text{OSiMe}_3)$ ([link](#)a). However, it was found that except for deposition at very high temperatures (> 900 °C) the deposited films this mono-siloxide compound were aluminum-rich ($\text{Al}/\text{Si} = 1.3 - 2.1$) and thus showed thermal instability in the insulating properties caused by crystallization in the films. It would appear that in order for silicon-rich alumina-silica films to be grown more siloxane substituents are required, e.g., the *tris*-siloxo aluminum complex $[\text{Al}(\text{OSiEt}_3)_3]_2$ ([link](#)b).



Precursors for aluminum silicate thin films.

The Al/Si ratio of thin films growth by APCVD using $[\text{Al}(\text{OSiEt}_3)_3]_2$ at 420 - 550 °C, was found to be dependent on the deposition temperature and the carrier gas composition (O_2/Ar). This temperature and oxygen-dependent variation in the film composition suggests that two competing precursor decomposition pathways are present.

1. Deposition in the absence of O_2 , is similar to that observed for the decomposition of $\text{Al}(\text{O}^i\text{Pr})_2(\text{OSiMe}_3)$ under N_2 , and would imply that the film composition is determined by the temperature-dependent tendencies of the Al-O-Si bonds to cleave.
2. The temperature-independent oxidative decomposition of the precursor. While it is possible to prepare films richer in Si using $[\text{Al}(\text{OSiEt}_3)_3]_2$ rather than $\text{Al}(\text{O}^i\text{Pr})_2(\text{OSiMe}_3)$, the Al:Si ratio is unfortunately not easily controlled simply by the number of siloxy ligands per aluminum in the precursor.

Films grown from the single-source precursor $\text{Al}(\text{O}^i\text{Pr})_2(\text{OSiMe}_3)$ crystallize to kyanite, Al_2SiO_5 , whereas those grown from $[\text{Al}(\text{OSiEt}_3)_3]_2$ remained amorphous even after annealing.

Bibliography

- A. W. Apblett, L. K. Cheatham, and A. R. Barron, *J. Mater. Chem.*, 1991, **1**, 143.
- K. M. Gustin and R. G. Gordon, *J. Electronic Mater.*, 1988, **17**, 509.
- C. Landry, L. K. Cheatham, A. N. MacInnes, and A. R. Barron, *Adv. Mater. Optics Electron.*, 1992, **1**, 3.
- Y. Nakaido and S. Toyoshima, *J. Electrochem. Soc.*, 1968, **115**, 1094.
- T. Maruyama and T. Nakai, *Appl. Phys. Lett.*, 1991, **58**, 2079.
- K. Sawada, M. Ishida, T. Nakamura, and N. Ohtake, *Appl. Phys. Lett.*, 1988, **52**, 1673.
- J. Von Hoene, R. G. Charles, and W. M. Hickam, *J. Phys. Chem.*, 1958, **62**, 1098.

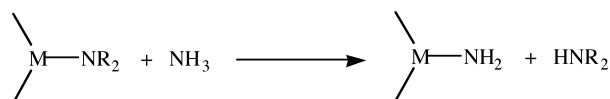
Introduction to Nitride Chemical Vapor Deposition

The refractory nature and high dielectric properties of many nitrides make them attractive for chemical and electronic passivation. As a consequence silicon nitride has become the standard within the semiconductor industry, as both an encapsulation layer and as an etch mask.

In a similar manner to oxide growth by chemical vapor deposition (CVD), two sources are generally required for binary nitride CVD: the element of choice and a nitrogen source. However, unlike the CVD of oxides, elemental nitrogen (N_2) is not reactive, even at elevated temperatures, thereby requiring plasma enhancement. Even with plasma enhanced CVD (PECVD), N_2 does not yield high quality films. As a substitute for N_2 , ammonia (NH_3) has found general acceptance as a suitable nitrogen source. It is a gas, readily purified and cheap, however, it is of low reactivity at low temperatures. PECVD has therefore found favor for low temperature NH_3 -based precursor systems.

Recent attempts to lower deposition temperatures have included the use of more reactive sources (e.g., H_2NNH_2) and precursors containing nitrogen as a coordinated ligand. Probably the most important discovery with respect to nitride deposition is the use of a transamination reaction between amido compounds and ammonia ([link](#)).

Equation:



Bibliography

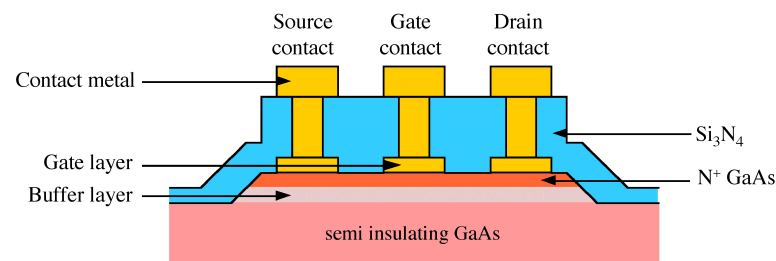
- D. M. Hoffman, *Polyhedron*, 1994, **13**, 1169.

Chemical Vapor Deposition of Silicon Nitride and Oxynitride

Introduction

Stoichiometric silicon nitride (Si_3N_4) is used for chemical passivation and encapsulation of silicon bipolar and metal oxide semiconductor (MOS) devices, because of its extremely good barrier properties for water and sodium ion diffusion. Water causes device metallization to corrode, and sodium causes devices to become electrically unstable. Silicon nitride is also used as a mask for the selective oxidation of silicon, and as a strong dielectric in MNOS (metal-nitride-oxide-silicon) structures.

The use of ion implantation for the formation of active layers in GaAs MESFET devices ([\[link\]](#)) allow for control of the active layer thickness and doping density. Since implantation causes structural disorder, the crystal lattice of the GaAs must be subjected to a post implantation rapid thermal anneal step to repair the damage and to activate the implanted species. The required annealing temperature ($> 800\text{ }^\circ\text{C}$) is higher than the temperature at which GaAs decomposes. Silicon nitride encapsulation is used to prevent such dissociation. Silicon nitride is also used for the final encapsulation of GaAs MESFET devices ([\[link\]](#)).



Schematic diagrams of a GaAs metal-semiconductor field effect transistor (MESFET). Adapted from A. R. Barron, in *CVD of Nonmetals*, Ed. W. S. Rees, Jr., Wiley, NY (1996).

The deposition of Si_3N_4 is a broadly practiced industrial process using either grown by low pressure CVD (LPCVD) or plasma enhanced CVD (PECVD) with comparable properties for the grown films ([\[link\]](#)).

Deposition	LPCVD	PECVD

Growth temperature (°C)	700 - 800	250 - 350
Composition	Si ₃ N ₄ (H)	SiN _x H _y
Si/N ratio	0.75	0.8 - 1.2
Atom% H	4 - 8	20 - 25
Dielectric constant	6 - 7	6 - 9
Refractive index	2.01	1.8 - 2.5
Resistivity (Ω.cm)	1016	106 - 1015
Band gap (eV)	5	4 - 5

Summary of the properties of silicon nitride grown in typical commercial systems.

One of the disadvantages of Si₃N₄ is its high dielectric constant that may limit device speed at higher operating frequencies. It is hoped that silicon oxynitride (SiON) films will exhibit the best properties of Si₃N₄ and SiO₂, namely the passivation and mechanical properties of Si₃N₄ and the low dielectric constant and low stress of SiO₂.

A summary of some typical CVD systems for silicon nitride is given in [\[link\]](#).

Silicon precursor	Nitrogen source	Carrier gas	CVD method	Deposition temp. (°C)	Comment
SiH ₄	NH ₃	N ₂	APCVD	70 - 900	
SiH ₄	NH ₃	Ar/N ₂	PECVD	20 - 600	Commercial process
SiH ₄	N ₂	N ₂	PECVD	70 - 300	Porous films
SiCl ₂ H ₂	NH ₃	N ₂	LPCVD	700 - 900	Commercial process
Si ₂ Cl ₆	NH ₃	-	LPCVD	450 - 850	
Et ₂ SiH ₂	NH ₃	-	LPCVD	650 - 725	C impurities
RSi(N ₃) ₃ (R	-	-	LPCVD	450 - 600	Danger –

= Et, ^t Bu)					precursor explosive
MeSiH(NH) _n	-	NH ₃ /H ₂	APCVD	600 - 800	Significant C content
Si(NMe ₂) ₄ - nH _n	-	He	APCVD	600 - 750	Significant C content
Si(NMe ₂) ₄ - nH _n	NH ₃	He	APCVD	600 - 750	No C contamination

Precursors and deposition conditions for Si₃N₄ CVD.

CVD of silicon nitride from hydrides and chlorides

The first commercial growth of silicon nitride was by the reaction of SiH₄ and NH₃ by either atmospheric pressure CVD (APCVD) or PECVD. Film growth using APCVD is slower and requires higher temperatures and so it has been generally supplanted by plasma growth, however, film quality for APCVD is higher due to the lower hydrogen content. While thermally grown films are close to stoichiometric, PECVD films have a composition in which the S/N ratio is observed to vary from 0.7 - 1.1. The non-stoichiometric nature of PECVD films is explained by the incorporation of significant hydrogen in the films (10 - 30%). PECVD of SiN_x using SiH₄/N₂ leads to electronically leaky films due to the porous nature of the films, however, if an electron cyclotron resonance (ECR) plasma is employed, SiN_x films of high quality may be deposited on ambient temperature substrates.

The more recent commercial methods for silicon nitride deposition involves LPCVD using SiCl₂H₂ as the silicon source in combination with NH₃ at 700 - 900 °C. The reduced pressure of LPCVD has the advantages of high purity, low hydrogen content, stoichiometric films, with a high degree of uniformity, and a high wafer throughput. It is for these reasons that LPCVD is now the method of choice in commercial systems. A large excess of NH₃ is therefore used in commercial systems to obtain stoichiometric films. Silicon nitride has also been prepared from SiCl₄/NH₄, SiBr₄/NH₃, and, more recently, Si₂Cl₆/NH₃.

Silicon oxynitride (SiON) may be prepared by the use of any of the precursors used for silicon nitride with the addition of either N₂O or NO as an oxygen source. The composition and properties of the SiO_xN_y films may be varied from SiO₂-like to Si₃N₄-like by the variation of the reactant flow rates.

SiCl₂H₂ gas plumbing to a LPCVD reactor must be thermally insulated to prevent condensation that would otherwise lead to hazy deposits on the film. The volatile by-products from CVD produce NH₄Cl at the exhaust of the reaction tube, and in the plumbing and pumping system. It would be desirable, therefore, to find an alternative, chlorine-free silicon source with none of the toxicity or pyrophoricity problems associated with SiH₄. It is for this reason that organosilicon compounds have been investigated.

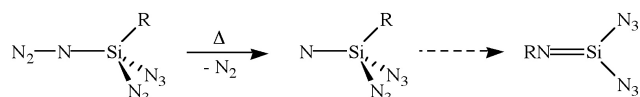
CVD from organosilicon precursors

Diethylsilane, Et_2SiH_2 , has shown promise as a replacement for SiH_4 in the low temperature LPCVD of SiO_2 , and has been investigated as a source for SiN_x and SiO_xN_y films. Deposition by LPCVD in the presence of NH_3 produces SiN_x films, in which the carbon contamination (4 - 9%) depends on the partial pressure of the Et_2SiH_2 . The presence of carbon raises the refractive index (2.025 - 2.28) with respect to traditional LPCVD films (2.01). Mixtures of Et_2SiH_2 , NH_3 , and N_2O deposit SiO_xN_y films where the composition is controlled by the $\text{NH}_3:\text{N}_2\text{O}$ ratio.

CVD from silicon-nitrogen compounds

The incorporation of carbon into silicon nitride films is a persistent problem of organosilicon precursors. Several studies have been aimed at developing single source precursors containing a Si-N bond rather than Si-C bonds. Polyazidosilanes, $\text{R}_n\text{Si}(\text{N}_3)_{4-n}$, are low in carbon and hydrogen, reasonably volatile, and contain highly activated nitrogen, however, they represent a significant explosive hazard: *they are explosive with an equivalent force to TNT*. Films deposited using $\text{EtSi}(\text{N}_3)_3$ and $(\text{tBu})\text{Si}(\text{N}_3)_3$ showed promise, despite the observation of oxygen and carbon. Pyrolytic studies on the azide precursors suggest that the primary decomposition step is the loss of dinitrogen, which is followed by migration of the alkyl onto the remaining nitrogen, [\[link\]](#). The fact that neither the addition of NH_3 or H_2 influence the film deposition rate suggest that the intramolecular nitride formation process is fast, relative to reaction with NH_3 , or hydrogenation.

Equation:



Carbon incorporation is also observed for the APCVD deposition from $\text{Si}(\text{NMe}_2)_n\text{H}_{4-n}$ ($n = 2 - 4$). However, using the Hoffman transamination reaction, deposition in the presence of NH_3 completely removed carbon incorporation into the stoichiometric Si_3N_4 film. From FTIR data, the hydrogen content was estimated to be 8 - 10 atom percent. While the $\text{Si}(\text{NMe}_2)_n\text{H}_{4-n}/\text{NH}_3$ system does not provide substantially lower temperatures than APCVD using SiH_4/NH_3 growth rates are significantly higher. Unlike the azide precursors, $\text{Si}(\text{NMe}_2)_n\text{H}_{4-n}$ are easier to handle than either SiH_4 or SiCl_2H_2 .

Bibliography

- J. C. Barbour, H. J. Stein, O. A. Popov, M. Yoder, and C.A. Outten, *J. Vac. Sci. Technol. A.*, 1991, **9**, 480.
- J. A. Higgins, R. L. Kuvas, F. H. Eisen, and D. R. Chen, *IEEE Trans. Electron. Devices*, 1978, **25**, 587.
- D. M. Hoffman, *Polyhedron*, 1994, **13**, 1169.
- W. Kellner, H. Kniepkamp, D. Repow, M. Heinzl, and H. Boroleka, *Solid State Electron.*, 1977, **20**, 459.

- T. Makino, *J. Electrochem. Soc.*, 1983, **130**, 450.
- C. T. Naber and G. C. Lockwood, in *Semiconductor Silicon*, Eds. H. R. Huff and R. R. Burgess. The Electrochemical Society, Softbound Proceedings Series, Princeton, NJ (1973).
- J. E. Schoenholtz, D. W. Hess, *Thin Solid Films*, 1987, **148**, 285.

Chemical Vapor Deposition of Aluminum Nitride

Introduction

Aluminum nitride (AlN) has potential for significant applications in microelectronic and optical devices. It has a large direct bandgap ($E_{g,dir} = 6.28$ eV), extremely high melting point (3000 °C), high thermal conductivity (2.6 W/cm.K), and a large dielectric constant ($\epsilon = 9.14$). In present commercial microelectronic devices, AlN is used most often as a packaging material, allowing for the construction of complex packages with many signal, ground, power, bonding, and sealing layers. Aluminum nitride is especially useful for high power applications due to its enhanced thermal conductivity. Chemical vapor deposition (CVD) grown thin films of AlN have been centered upon its use as a high gate-insulation layer for MIS devices, and a dielectric in high-performance capacitors. One additional property of AlN that makes it a promising insulating material for both Si and GaAs devices is that its thermal expansion coefficient is almost identical to both of these semiconductors.

The lack of a suitably volatile homoleptic hydride for aluminum (AlH_3 is an involatile polymeric species) led to the application of aluminum halides and organometallic compounds as precursors. A summary of selected precursor combinations is given in [\[link\]](#).

Aluminum precursor	Nitrogen source	Carrier gas	CVD method	Deposition temp. (°C)	Comments
$AlCl_3(NH_3)$	-	N_2	LPCVD	700 - 1400	NH_4Cl present
$AlBr_3$	NH_3	N_2	APCVD	400 - 900	Br present
$AlBr_3$	N_2	N_2	LPCVD	520 - 560	oriented growth
$AlMe_3$	NH_3	H_2	LPCVD	1200	
$AlMe_3$	NH_3	He	APCVD	350 - 400	
$AlMe_3$	pre-cracked NH_3	H_2/He	APCVD	310 - 460	N-H and AlN-N bonds detected
$AlMe_3$	$tBuNH_2$	H_2	APCVD	400 - 600	high C

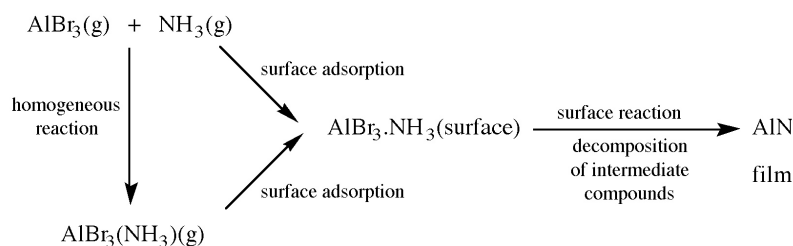
	or $i\text{PrNH}_2$				content, low N
AlMe_3	Me_3SiN_3	H_2	APCVD	300 - 450	very high C content
$[\text{R}_2\text{Al}(\text{NH}_2)]_3$ (R = Me, Et)		H_2	LPCVD	400 - 800	poor film quality, high C content
$[\text{R}_2\text{AlN}_3]_3$ (R = Me, Et)		-	LPCVD	400 - 500	unreacted precursor present on film
$\text{Al}(\text{NMe}_2)_3$	NH_3	He	APCVD	100 - 500	amorphous 100 - 200 °C, crystalline 300 - 500 °C

Precursors and deposition conditions for AlN CVD.

CVD from halides

The observation that AlN powder may be produced upon the thermal decomposition of the $\text{AlCl}_3(\text{NH}_3)$ complex, prompted initial studies on the use of $\text{AlCl}_3/\text{NH}_3$ for the CVD of AlN films. Initially, the low volatility of AlCl_3 (a polymeric chain structure) required that the $\text{AlCl}_3(\text{NH}_3)$ complex to be used as a single precursor. Low pressure CVD (LPCVD) at 5 -10 Torr resulted in deposition of AlN films, although films deposited below 1000 °C were contaminated with NH_4Cl , and all the films contained chlorine. Films with reasonable electrical properties were prepared by the use of the more volatile *tris*-ammonia complex, $\text{AlCl}_3(\text{NH}_3)_3$. The dielectric constant for films grown at 800 - 1000 °C (11.5) is higher than bulk AlN (9.14) and also than that of the films grown at 1100 °C (8.1). All the films were polycrystalline with the grain size increasing with increasing deposition temperatures and preferred orientation was observed only for the films grown below 1000 °C.

Aluminum bromide is a dimeric volatile compound, $[\text{Br}_2\text{Al}(\mu\text{-Br})]_2$, and is more attractive as a CVD source, than AlCl_3 . Deposition of AlN films can be accomplished using AlBr_3 and NH_3 in an APCVD system with H_2 as the carrier gas. The mechanism of film growth has been proposed ([\[link\]](#)).



Mechanism of APCVD film growth of AlN using AlBr₃ and NH₃.

Due to the high temperatures required (750 °C) for good quality AlN film growth from AlBr₃, PECVD was investigated. Using an AlBr₃-H₂-N₂ gas mixture and a 2450 MHz microwave (100 - 1000 W) plasma source, AlN films were grown. The maximum deposition rate occurred with an N₂/AlBr₃ ratio of *ca.* 20 and a substrate temperature *ca.* 600 °C.

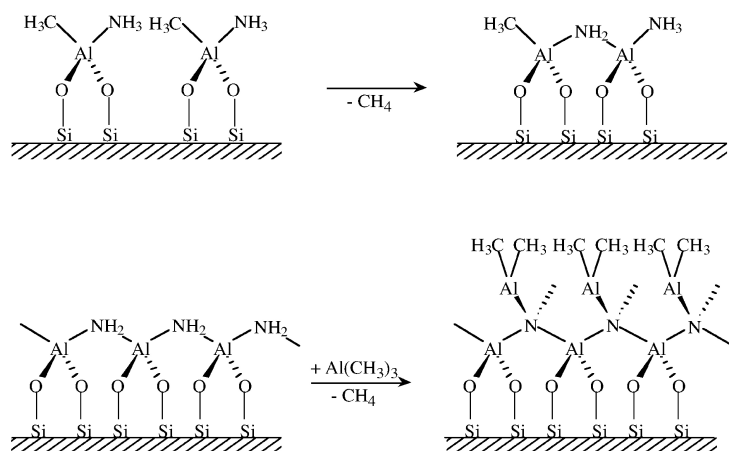
CVD from aluminum alkyls

Based upon the successful metal organic CVD (MOCVD) growth of AlGaAs using the alkyl derivatives, AlR₃, it was logical to extend MOCVD to aluminum nitride. Initial studies were performed using AlMe₃ and NH₃ with H₂ carrier gas. While these films are generally of high quality, the temperature of deposition is incompatible with semiconductor processing (being above both the melting point of most metallization alloys and the temperature at which dopant migration becomes deleterious). Lower temperatures (as low as 350 °C) were explored, however significant pre-reaction was observed between AlMe₃ and NH₃; causing depletion of the reactants in the deposition zone, reducing the growth rate and leading to non-uniform deposits. Two routes have been investigated by which this problem can be circumvented.

PECVD successfully lowers the deposition temperature, although, degradation of the substrate surface by ion bombardment is a significant drawback. Given that it is the ammonia decomposition that represents the highest energy process, pre-cracking should lower the overall deposition temperature. This is indeed observed for the AlMe₃/NH₃-based AlN system where growth is achieved as low as 584 °C if the NH₃ is catalytically cracked over a heated tungsten filament (1747 °C). In fact, with catalytic pre-cracking, deposition rates were observed to be an order of magnitude greater than for PECVD at the same temperatures, resulting in films that were crystalline with columnar growth. For this approach to low-temperature MOCVD growth of AlN the only major drawback is the presence of residual N-H and AlN-N groups detected by FT-IR.

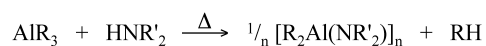
Chemical solutions to the high stability of NH₃ have primarily centered upon the use of alternative nitrogen sources. The use of the volatile nitrogen source hydrazine (N₂H₄), has allowed for the growth of AlN at temperatures as low as 220 °C, however, hydrazine is extremely toxic and highly unstable, restricting its commercial application. Primary amines, such as ^tBuNH₂ or ⁱPrNH₂, allow for deposition at modest temperatures (400 - 600 °C). The

Interest in the mechanism of nucleation and atomic layer growth of AlN has prompted several mechanistic studies of the formation of Al-N bonds on the growth surface. All the studies concurred that the mechanism involves a step-wise reaction where the amide (-NH₂-) groups form covalent bonds to aluminum irrespective of substrate. A schematic representation of the process is shown in [\[link\]](#).

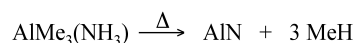


CVD from aluminum amide and related compounds

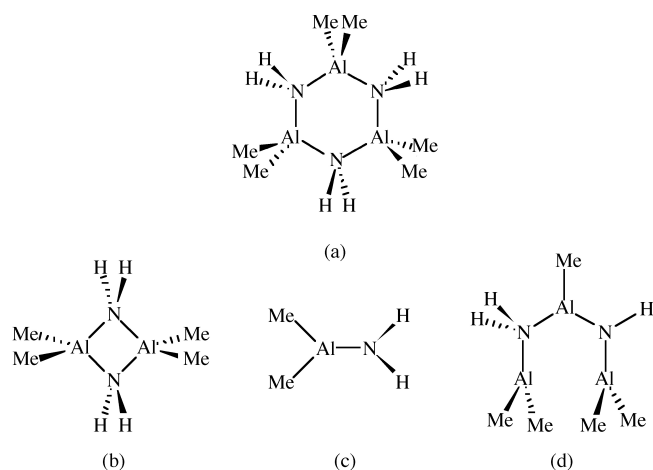
Equation:



Equation:



The trimeric dimethylaluminum amide, $[\text{Me}_2\text{Al}(\text{NH}_2)]_3$ ([link](#)a), was originally used as a single source precursor for growth of AlN under LPCVD conditions using a hot walled reactor, although subsequent deposition was also demonstrated in a cold walled system. Film quality was never demonstrated for electronic applications, but the films showed promise as fiber coatings for composites. The concept of using a trimeric single source precursor for AlN was derived from the observation of Al_3N_3 cycles as the smallest structural fragment in wurtzite AlN. However, detailed mechanistic studies indicate that under gas phase thermolysis the trimeric precursor $[\text{Me}_2\text{Al}(\text{NH}_2)]_3$ is in equilibrium with (or decomposes to) dimeric ([link](#)b) and monomeric ([link](#)c) compounds. Furthermore, nitrogen-poor species ([link](#)d) were also observed by TOF-mass spectrometry.



The trimeric dimethylaluminum amide (a) used as a single source precursor for growth of AlN, and the decomposition products (b - d) observed by TOF-mass spectrometry.

Following the early reports of single source precursor routes, a wide range of compounds have been investigated, including $[\text{Al}(\text{NR}_2)_3]_2$, $[\text{HAl}(\text{NR}_2)_2]_2$ ($\text{R} = \text{Me}, \text{Et}$), and $[\text{Me}_2\text{AlN}(\text{iPr})_2]_2$, all of which gave AlN, but none of these precursors give films of superior quality comparable to that obtained from traditional CVD. In particular, the films contained significant carbon contamination, prompting further investigations into the efficacy of, N-C bond free, dialkylaluminum azides, $[\text{R}_2\text{Al}(\text{N}_3)]_3$, as LPCVD precursors.

While aluminum *tris*-amides, $\text{Al}(\text{NR}_2)_3$ were shown to give carbon-contaminated films, APCVD carried-out with NH_3 as the carrier gas results in carbon-free AlN film growth as low as 100°C . The reason for the deposition of high quality films at such low temperatures resides

with the Hoffman transamination reaction between the primary amido unit and ammonia. The crystallinity, bandgap and refractive index for the AlN grown by APCVD using $[\text{Al}(\text{NMe}_2)_3]_2$ and NH_3 are dependent on the deposition temperature. Films grown at 100 - 200 °C are amorphous and have a low bandgap and low refractive index. Above 300 °C, the films are crystalline, and have a refractive index close to that of bulk AlN (1.99 - 2.02), with a bandgap (≤ 5.77 eV) approaching the values reported for polycrystalline AlN (5.8 - 5.9 eV).

Bibliography

- J. L. Dupuie and E. Gulari, *J. Vac. Sci. Technol. A*, 1992, **10**, 18.
- D. M. Hoffman, *Polyhedron*, 1994, **13**, 1169.
- L. V. Interrante, W. Lee, M. McConnell, N. Lewis, and E. Hall, *J. Electrochem. Soc.*, 1989, **136**, 472.
- H. M. Manasevit, F. M. Erdmann, and W. I. Simpson, *J. Electrochem. Soc.*, 1971, **118**, 1864.
- Y. Pauleau, A. Bouteville, J.J. Hantzpergue, J. C Remy, and A. Cachard, *J. Electrochem. Soc.*, 1980, **127**, 1532.
- Y. Someno, M. Sasaki, and T. Hirai, *Jpn. J. Appl. Phys.*, 1990, **29**, L358.

Metal Organic Chemical Vapor Deposition of Calcium Fluoride

The chemical vapor deposition (CVD) of metal fluorides has been much less studied than that of oxides, pnictides, or chalcogenides. As may be expected where a volatile fluoride precursor is available then suitable films may be grown. For example, Group 5 (V, Nb, Ta), 6 (Mo, W), and 7 (Re) transition metals are readily deposited from fluoride-hydrogen mixtures. While the use of fluorine is discouraged on safety grounds, many of the fluorinated alkoxide or β -diketonate ligands employed for metal oxide metal organic chemical vapor deposition (MOCVD) are predisposed to depositing metal fluorides. The use of fluorine substituted derivatives is because they are often more volatile than their hydrocarbon analogs, and therefore readily used for both atmospheric and low pressure CVD. To minimize the unwanted formation of metal fluorides, water vapor is incorporated in the gas stream, and it is common to perform post-deposition hydrolytic anneals. However, there exist a number of applications where fluorides are required. For example, the highly insulating nature of CaF_2 and SrF_2 has prompted investigations into their use as a gate insulator in GaAs-based metal insulator semiconductor field effect transistor (MISFET) devices. It should be noted that while CaF_2 is a good insulator, the CaF_2/GaAs interface has a high interface trap density, requiring a passivation buffer layer to be deposited on GaAs prior to CaF_2 growth.

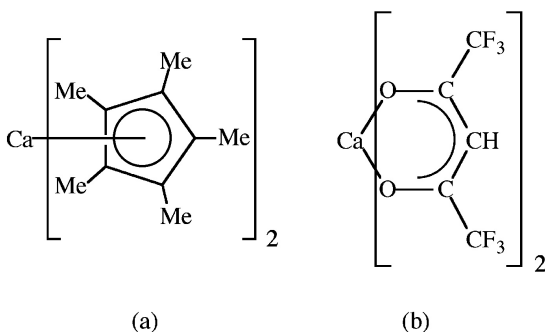
One of the difficulties with the use of CaF_2 (and SrF_2) on GaAs is the lattice mismatch ([\[link\]](#)), but this may be minimized by the use of solid solutions between CaF_2 - SrF_2 . The composition $\text{Ca}_{0.44}\text{Sr}_{0.56}\text{F}_2$ is almost perfectly lattice-matched to GaAs. Unfortunately, the thermal expansion coefficient differences between GaAs and CaF_2 - SrF_2 produce strains at the film/substrate interface under high temperature growth conditions. The solution to this latter problem lies in the low temperature deposition of CaF_2 - SrF_2 by CVD.

Compound	Lattice constant (Å)
----------	----------------------

CaF ₂	5.46
SrF ₂	5.86
BaF ₂	6.20
GaAs	5.6532

Lattice parameters of Group 2 (II) fluorides in comparison with GaAs.

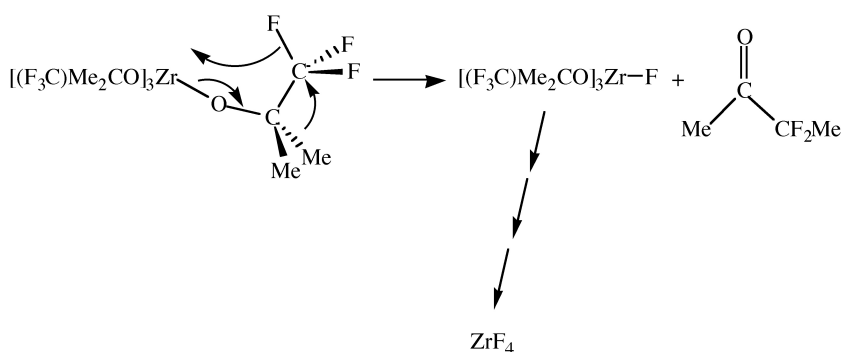
Polycrystalline CaF₂ may be grown by the pyrolytic decomposition of Ca(C₅Me₅)₂ ([link](#)a) in either SiF₄ or NF₃. Deposition at 150 °C results in polycrystalline films with high levels of carbon (18%) and oxygen (7%) impurities limiting the films usefulness in electronic applications. However, significantly higher purity films may be grown at 100 °C using the photo-assisted decomposition of Ca(hfac)₂ ([link](#)b). These films were deposited at 30 Å/min and showed a high degree of crystallographic preferred orientation.



CaF₂ MOCVD precursors.

The mechanism enabling fluoride transfer to the metal (from the carbon of fluorinated alkoxide ligands) has been investigated. MOCVD employing [Na(OR_f)]₄ and Zr(OR_f)₄ [OR_f = OCH(CF₃)₂ and OCMe_{3-n}(CF₃)_n, n = 1 - 3] gives NaF and ZrF₄ films, respectively, with volatile fluorocarbon side-products. Analysis of the organic side-products indicated that

decomposition occurs by transfer of fluorine to the metal in conjunction with a 1,2-migration of a residual group on the alkoxide, to form a ketone ([link](#)). The migration is increasingly facile in the order $\text{CF}_3 \ll \text{CH}_3 \leq \text{H}$. The initial M-F bond formation has been proposed to be as a consequence of the close M...F agostic interactions observed for some fluoroalkoxide and fluoro- β -diketonates.



Proposed mechanism for the decomposition of fluorinated alkoxide compounds. (Adapted from J. A. Samuels, W. -C. Chiang, C. -P. Yu, E. Apen, D. C. Smith, D. V. Baxter, K. G. Caulton, *Chem. Mater.*, 1994, **6**, 1684).

Bibliography

- A. R. Barron, in *CVD of Nonmetals*, W. S. Rees, Jr. (ed), Wiley, New York (1996).
- B. D. Fahlman and A. R. Barron, *Adv. Mater. Opt. Electron.*, 2000, **10**, 223.
- H. Heral, L. Bernard, A. Rocher, C. Fontaine, A. Munoz-Jague, *J. Appl. Phys.*, 1987, **61**, 2410.
- J. A. Samuels, W. -C. Chiang, C. -P. Yu, E. Apen, D. C. Smith, D. V. Baxter, K. G. Caulton, *Chem. Mater.*, 1994, **6**, 1684.
- W. Vere, K. J. Mackey, D. C. Rodway, P. C. Smith, D. M Frigo, D. C. Bradley, *Angew. Chem. Int. Ed. Engl. Adv. Mater.*, 1989, **28**, 1581.

Precursors for Chemical Vapor Deposition of Copper

Note: This module was developed as part of the Rice University course CHEM-496: *Chemistry of Electronic Materials*. This module was prepared with the assistance of Wei Zhao.

Introduction

Chemical vapor deposition (CVD) is a process for depositing solid elements and compounds by reactions of gas-phase molecular precursors. Deposition of a majority of the solid elements and a large and ever-growing number of compounds is possible by CVD.

Most metallization for microelectronics today is performed by the physical vapor deposition (PVD) processes of evaporation and sputtering, which are often conceptually and experimentally more straightforward than CVD. However, the increasing importance of CVD is due to a large degree to the advantages that it holds over physical vapor deposition. Foremost among these are the advantages of conformal coverage and selectivity. Sputtering and evaporation are by their nature line-of-sight deposition processes in which the substrate to be coated must be placed directly in front of the PVD source. In contrast, CVD allows any substrate to be coated that is in a region of sufficient precursor partial pressure. This allows the uniform coating of several substrate wafers at once, of both sides of a substrate wafer, or of a substrate of large size and/or complex shape. The PVD techniques clearly will also deposit metal on any surface that is in line of sight. On the other hand, it is possible to deposit selectively on some substrate materials in the presence of others using CVD, because the deposition is controlled by the surface chemistry of the precursor/substrate pair. Thus, it may be possible, for example, to synthesize a CVD precursor that under certain conditions will deposit on metals but not on an insulating material such as SiO_2 , and to exploit this selectivity, for example, in the fabrication of a very large-scale integrated (VLSI) circuit. It should also be pointed out that, unlike some PVD applications, CVD does not cause radiation damage of the substrate.

Since the 1960s, there has been considerable interest in the application of metal CVD for thin-film deposition for metallization of integrated circuits. Research on the thermal CVD of copper is motivated by the fact that copper has physical properties that may make it superior to either tungsten or aluminum in certain microelectronics applications. The resistivity of copper (1.67 mW.cm) is much lower than that of tungsten (5.6 mW.cm) and significantly lower than that of aluminum (2.7 mW.cm). This immediately suggests that copper could be a superior material for making metal interconnects, especially in devices where relatively long interconnects are required. The electromigration resistance of copper is higher than that of aluminum by four orders of magnitude. Copper has increased resistance to stress-induced voidage due to its higher melting point versus aluminum. There are also reported advantages for copper related device performance such as greater speed and reduced cross talk and smaller RC time constants. On the whole, the combination of superior resistivity and intermediate reliability properties makes copper a promising material for many applications, provide that suitable CVD processes can be devised.

Applications of metal CVD

There are a number of potential microelectronic applications for metal CVD, including gate metallization (deposit on semiconductor), contact metallization (deposit on semiconductor), diffusion barrier metallization (deposit on semiconductor), interconnect metallization (deposit on insulator and conductor or semiconductor). Most of the relevant features of metal CVD are found in the interconnect and via fill applications, which we briefly describe here. There are basically two types of metal CVD processes that may occur:

1. Blanket or nonselective deposition, in which deposition proceeds uniformly over a variety of surfaces.
2. Selective deposition in which deposition only occurs on certain types of surfaces (usually semiconductors or conductors, but not insulators).

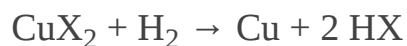
A primary application of blanket metal CVD is for interconnects. The conformal nature of the CVD process is one of the key advantages of CVD over PVD and is a driving force for its research and development. The degree of conformality is usually described as the “step coverage”, which is normally defined as the ratio of the deposit thickness on the step sidewall to the deposit thickness on the top surface. Another application for blanket metal CVD is via hole filling to

planarize each level for subsequent processing, This is achieved by depositing a conformal film and etching back to the insulator surface, leaving the metal “plug” intact. Another unique aspect of CVD is its potential to deposit films selectively, which would eliminate several processing steps required to perform the same task. The primary application for selective metal CVD would be for via hole filling. Ideally, deposition only occurs on exposed conductor or semiconductor surfaces, so filling of the via hole is achieved in a single step.

Copper CVD

The chemical vapor deposition of copper originally suffered from a lack of readily available copper compounds with the requisite properties to serve as CVD precursors. The successful development of a technologically useful copper CVD process requires first and foremost the design and synthesis of a copper precursor which is volatile, i.e., possesses an appreciable vapor pressure and vaporization rate to allow ease in transportation to the reaction zone and deposition at high growth rates. Its decomposition mechanism(s) should preferably be straightforward and lead to the formation of pure copper and volatile by-products that are nonreactive and can be cleanly removed from the reaction zone to prevent film, substrate, and reactor contamination. Gaseous or liquid sources are preferred to solid sources to avoid undesirable variations in vaporization rates because of surface-area changes during evaporation of solid sources and to permit high levels of reproducibility and control in source delivery. Other desirable features in precursor selection include chemical and thermal stability to allow extended shelf life and ease in transport and handling, relative safety to minimize the industrial and environmental impact of processing and disposal, and low synthesis and production costs to ensure an economically viable process.

Several classes of inorganic and metalorganic sources have been explored as copper sources. Inorganic precursors for copper CVD used hydrogen reduction of copper halide sources of the type CuX or CuX_2 , where X is chlorine (Cl) or fluorine (F):



The volatility of copper halides is low, the reactions involved require prohibitively high temperatures (400 - 1200 °C), lead to the production of corrosive by-products such as hydrochloric and hydrofluoric acids (HCl and HF), and produce deposits with large concentrations of halide contaminants. Meanwhile, the exploration of metalorganic chemistries has involved various copper(II) and copper(I) source precursors, with significant advantages over inorganic precursors.

From Cu(II) precursors

Volatile Cu(II) compounds

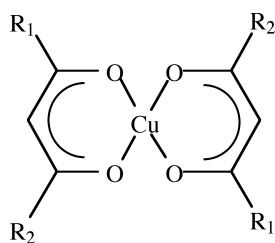
Copper was known to form very few stable, volatile alkyl or carbonyl compounds. This was thought to eliminate the two major classes of compounds used in most existing processes for CVD of metals or compound semiconductors. Copper halides have been used for chemical vapor transport growth of Cu-containing semiconductor crystals. But the evaporation temperatures needed for copper halides are much higher than those needed for metal-organic compounds. Film purity and resistivity were also a problem, possibly reflecting the high reactivity of Si substrates with metal halides.

Cu(II) compounds that have been studied as CVD precursors are listed in [\[link\]](#). The structural formulas of these compounds are shown in [\[link\]](#) along with the ligand abbreviations in [\[link\]](#). Each compound contains a central Cu(II) atom bonded to two singly charged β -diketonate or β -ketoiminate ligands. Most of them are stable, easy to synthesize, transport and handle.

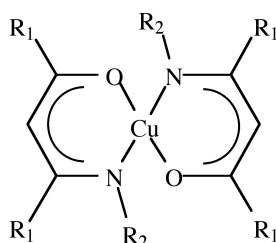
Compound	Evaporation temp. (°C)	Deposition temp. (°C)	Carrier gas	Reactor pressure (Torr)
Cu(acac) ₂	180 - 200	225 - 250	H ₂ /Ar	760
Cu(hfac) ₂	80 - 95	250 - 300	H ₂	760

Cu(tfac) ₂	135 - 160	250 - 300	H ₂	760
Cu(dpm) ₂	100	400	none	<10 ⁻²
Cu(ppm) ₂	100	400	none	<0.3
Cu(fod) ₂	-	300 - 400	H ₂	10 ⁻³ - 760
Cu(acim) ₂	287	400	H ₂	730
Cu(nona-F) ₂	85 - 105	270 - 350	H ₂	10 - 70
Cu(acen) ₂	204	450	H ₂	730

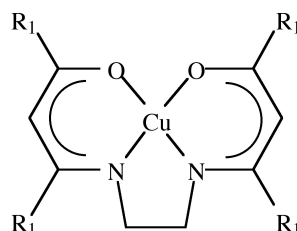
Studies of Cu CVD using Cu(II) compound. Adapted from T. Kodas and M. Hampden-Smith, *The Chemistry of Metal CVD*, VCH Publishers Inc., New York, NY (1994).



(a)



(b)



(c)

Structures of Cu(II) compounds studied as CVD precursors.

Ligand abbreviation	R ₁	R ₂	Structural type
acac	CH ₃	CH ₃	a
hfac	CF ₃	CF ₃	a
tfac	CH ₃	CF ₃	a
dpm	C(CH ₃) ₃	C(CH ₃) ₃	a
ppm	C(CH ₃) ₃	CF ₂ CF ₃	a
fod	C(CH ₃) ₃	CF ₂ CF ₂ CF ₃	a
acim	CH ₃	H	b
nona-F	CF ₃	CH ₂ CF ₃	b
acen	CH ₃	-	c

Ligand abbreviations for the structures shown in [\[link\]](#).

Attention has focused on Cu(II) β -diketonate [i.e., Cu(tfac)₂, Cu(hfac)₂] and Cu(II) β -ketoiminate [i.e., Cu(acim)₂, Cu(acen)₂]. An important characteristic of Cu(II) compounds as CVD precursors is the use of heavily fluorinated ligand such as Cu(tfac)₂ and Cu(hfac)₂ versus Cu(acac)₂. The main effort of fluorine substitution is a significant increase in the volatility of the complex.

Synthesis of Cu(II) precursors

Cu(hfac)₂·nH₂O (n = 0, 1, 2)

Cu(hfac)₂ is by far the most extensively studied of the Cu(II) CVD precursors. Preparations in aqueous solutions yield the yellow-green dihydrate, Cu(hfac)₂·2H₂O. This is stable in very humid air or at lower temperatures but

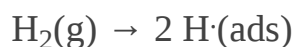
slowly loses one molecule of water under typical laboratory conditions to form the “grass-green” monohydrate, $\text{Cu}(\text{hfac})_2 \cdot \text{H}_2\text{O}$. The monohydrate, which is commercially available, can be sublimed unchanged and melts at 133 – 136 °C. More vigorous drying over concentrated H_2SO_4 produces the purple anhydrous compound $\text{Cu}(\text{hfac})_2$ (mp = 95 – 98 °C). The purple material is hygroscopic, converting readily into the monohydrate. Other β -diketonate $\text{Cu}(\text{II})$ complexes are prepared by the similar method.

Schiff-base complexes

Schiff-base complexes include $\text{Cu}(\text{acim})_2$, $\text{Cu}(\text{acen})$ and $\text{Cu}(\text{nona-F})_2$. The first two of these can be prepared by mixing $\text{Cu}(\text{NH}_3)_4^{2+}$ (aq) with the pure ligand and by adding freshly prepared solid $\text{Cu}(\text{OH})_2$ to a solution of the ligand in acetone. The synthesis of $\text{Cu}(\text{nona-F})_2$, on the other hand, involved two important developments: the introduction of the silyl enol ether route to the ligand and its conversion in-situ into the desired precursor. The new approach to the ligand was required because, in contrast to non-fluorinated β -diketonates, $\text{H}(\text{hfac})$ reacts with amines to produce salts.

Reaction mechanism

Starting from the experimental results, a list of possible steps for Cu CVD via H_2 reduction of $\text{Cu}(\text{II})$ compounds would include the followings, where removal of adsorbed ligand from the surface is believed to be the rate limiting step:



where L represents any of the singly charged β -diketonate or β -ketoiminate ligands described before. This mechanism gives a clear explanation of the importance of hydrogen being present: in the absence of hydrogen, HL cannot desorb cleanly into the gas phase and ligand will tend to decompose on the surface, resulting in impurity incorporation into the growing film. The mechanism is also supported by the observation that the deposition reaction is

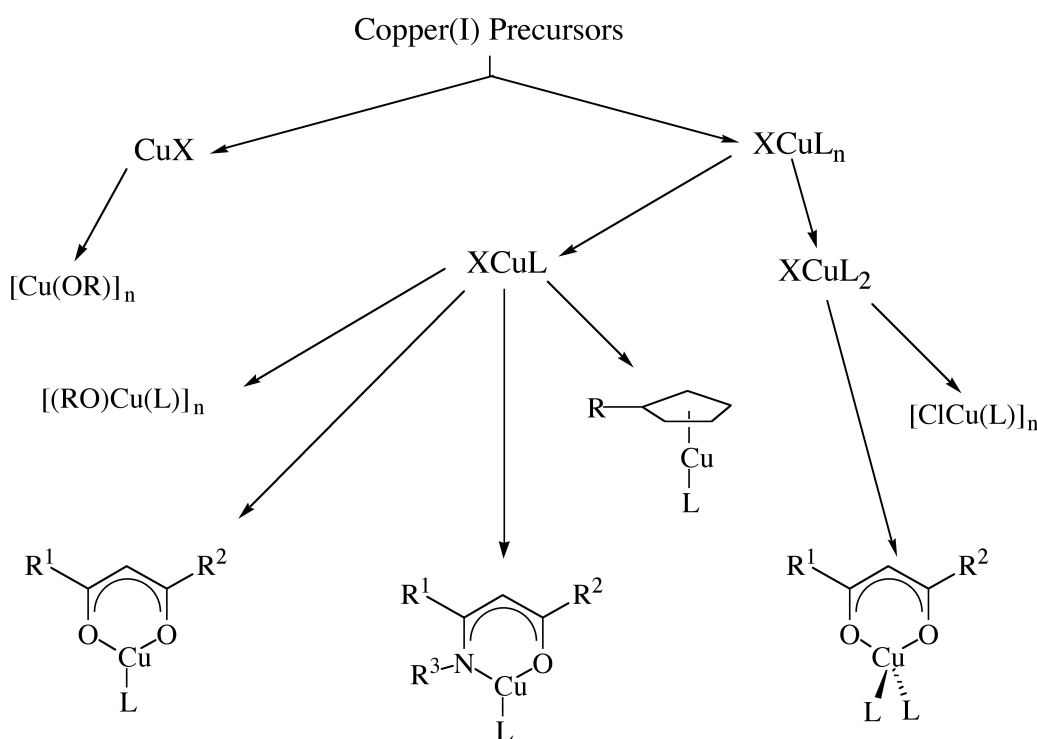
enhanced by the addition of alcohol containing β -hydrogen to the reaction mixture.

More recently, the focus has shifted to Cu(I) compounds including Cu(I) cyclopentadienyls and Cu(I) β -diketonate. The Cu(I) β -diketonate in particular show great promise as Cu CVD precursors and have superseded the Cu(II) β -diketonate as the best family of precursors currently available.

From Cu(I) precursors

Precursor design

The Cu(I) compounds that have been investigated are described in [\[link\]](#). These species can be broadly divided into two classes, CuX and XCuL_n , where X is a uninegative ligand and L is a neutral Lewis base electron pair donor. The XCuL_n class can be further subdivided according to the nature of X and L.



Copper(I) precursors used for CVD. Adapted from T. Kodas and M. Hampden-Smith, *The Chemistry of Metal CVD*, VCH

Publishers Inc., New York, NY (1994).

Compounds of general formula CuX are likely to be oligomeric resulting in a relatively low vapor pressure. The presence of a neutral donor ligand, L , is likely to reduce the extent of oligomerization compared to CuX by occupying vacant coordination sites. Metal alkoxide compounds are expected to undergo thermal decomposition by cleavage of either M-O or O-C bonds.

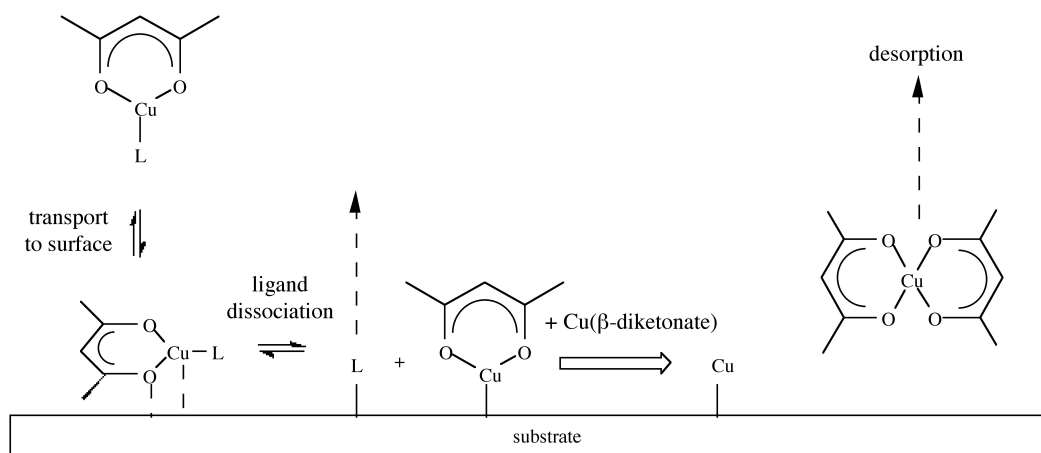
Organo-copper(I) compounds, RCuL , where R is alkyl, are thermally unstable, but cyclopentadienyl compounds are likely to be more robust due to the π -bonding of the cyclopentadienyl ligand to the copper center. At the same time, the cyclopentadienyl ligand is sterically demanding, occupies three coordination sites at the metal center, and thereby reduces the desire for oligomerization. In general, a cyclopentadienyl ligand is a poor choice to support CVD precursors, especially with electropositive metals, because this ligand is unlikely to be liable. Compounds in the family XCuL_2 , where X is a halide and L is a triorganophosphine, exhibit relatively high volatility but are thermally stable with respect to formation of copper at low temperatures. These species are therefore suitable as products of etching reactions of copper films.

A number of researchers have demonstrated the potential of a series of β -diketonate Cu(I) compounds, $(\beta\text{-diketonate})\text{CuL}_n$, where L is Lewis base and $n = 1$ or 2 , that fulfill most of the criteria outlined for precursor design before. These species were chosen as copper precursors for the following reasons:

- They contain the β -diketonate ligand which generally imparts volatility to metal-organic complexes, particularly when fluorinated, as a result of a reduction in hydrogen-bonding in the solid-state.
- They are capable of systematic substitution through both the β -diketonate and Lewis base ligands to tailor volatility and reactivity.
- Lewis bases such as phosphines, olefins and alkynes are unlikely to thermally decompose at temperatures where copper deposition occurs.
- These precursors can deposit copper via thermally induced disproportionation reactions and no ligand decomposition is required since the volatile Lewis base the Cu(II) disproportionation products are transported out of the reactor intact at the disproportionation temperature.

Reaction mechanism

A general feature of the reactions of Cu(I) precursors is that they thermally disproportionate, a mechanism likely to be responsible for the high purity of the copper films observed since ligand decomposition does not occur. The disproportionation mechanism is shown in [\[link\]](#) for (β -diketonate)CuL. The unique capabilities of this class of compounds result from this reaction mechanism by which they deposit copper. This mechanism is based on the dissociative adsorption of the precursor to form Cu(hfac) and L, disproportionation to form Cu(hfac)₂ and Cu and desorption of Cu(hfac)₂ and L.



Schematic diagram of the disproportionation mechanism. Adapted from T. Kodas and M. Hampden-Smith, *The Chemistry of Metal CVD*, VCH Publishers Inc., New York, NY (1994).

Thus, the starting material acts as its own reducing agent and no external reducing agent such as H₂ is required. Another advantage of the Cu(I) β -diketonates over the Cu(II) β -diketonates is that in the former the ligand L can be varied systematically, allowing the synthesis of a whole series of different but closely related compounds.

Selectivity

Selectivity deposition has been studied in both hot- and cold-wall CVD reactors as a function of the nature of the substrate, the temperature of the substrate and the nature of the copper substituents. Selectivity has usually been evaluated by using Si substrates on which SiO₂ has been grown and patterned with various metals by either electron-beam deposition, CVD or sputtering. Research has suggested that selectivity on metallic surfaces is attributable to the biomolecular disproportionation reaction involved in precursor decomposition.

Bibliography

- J. R. Creighton, and J. E. Parmeter, *Critical Review in Solid State and Materials Science*, 1993, **18**, 175.
- L. H. Dubois and B. R. Zegarski, *J. Electrochem. Soc.*, 1992, **139**, 3295.
- J. J. Jarvis, R. Pearce, and M. F. Lappert, *J. Chem. Soc., Dalton Trans.*, 1977, 999.
- A. E. Kaloyeros, A. Feng, J. Garhart, K. C. Brooks, S. K. Ghosh, A. N. Sazena, and F. Luehersch, *J. Electronic Mater.*, 1990, **19**, 271.
- T. Kodas and M. Hampden-Smith, *The Chemistry of Metal CVD*, VCH Publishers Inc., New York, NY (1994).
- C. F. Powell, J. H. Oxley, and J. M. Blocher Jr., *Vapor Deposition*, John Wiley, New York (1966).
- S. Shingubara, Y. Nakasaki, and H. Kaneko, *Appl. Phys. Lett.*, 1991, **58**, 42.

Rutherford Backscattering of Thin Films

Introduction

One of the main research interests of the semiconductor industry is to improve the performance of semiconducting devices and to construct new materials with reduced size or thickness that have potential application in transistors and microelectronic devices. However, the most significant challenge regarding thin film semiconductor materials is measurement. Properties such as the thickness, composition at the surface, and contamination, all are critical parameters of the thin films. To address these issues, we need an analytical technique which can measure accurately through the depth of the of the semiconductor surface without destruction of the material. Rutherford backscattering spectroscopy is a unique analysis method for this purpose. It can give us information regarding in-depth profiling in a non-destructive manner. However X-ray photo electron spectroscopy (XPS), energy dispersive X-ray analysis (EDX) and Auger electron spectroscopy are also able to study the depth-profile of semiconductor films. [\[link\]](#) demonstrates the comparison between those techniques with RBS.

Method	Destructive	Incident particle	Outgoing Particle	Detection limit	Depth resolution
RBS	No	Ion	Ion	~1	10 nm
XPS	Yes	X-ray photon	Electron	~0.1-1	~1 μm
EDX	Yes	Electron	X-ray photon	~0.1	1.5 nm
Auger	Yes	Electron	Electron	~0.1-1	1.5 nm

Comparison between different thin film analysis techniques.

Basic concept of Rutherford backscattering spectroscopy

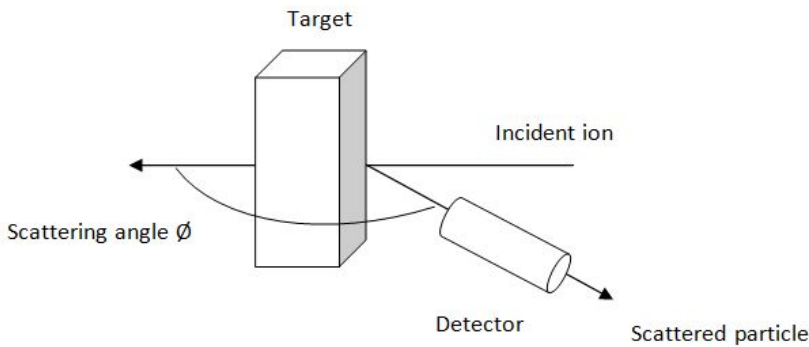
At a basic level, RBS demonstrates the electrostatic repulsion between high energy incident ions and target nuclei. The specimen under study is bombarded with monoenergetic beam of $^4\text{He}^+$ particles and the backscattered particles are detected by the detector-analysis system which measures the energies of the particles. During the collision, energy is transferred from

the incident particle to the target specimen atoms; the change in energy of the scattered particle depends on the masses of incoming and target atoms. For an incident particle of mass M_1 , the energy is E_0 while the mass of the target atom is M_2 . After the collision, the residual energy E of the particle scattered at angle θ can be expressed as:

$$E = k^2 E_0$$

$$k = \frac{\left(M_1 \cos \theta + \sqrt{M_2^2 - M_1^2 \sin^2 \theta} \right)}{M_1 + M_2}$$

where k is the kinematic scattering factor, which is actually the energy ratio of the particle before and after the collision. Since k depends on the masses of the incident particle and target atom and the scattering angle, the energy of the scattered particle is also determined by these three parameters. A simplified layout of backscattering experiment is shown in [Figure 1](#).



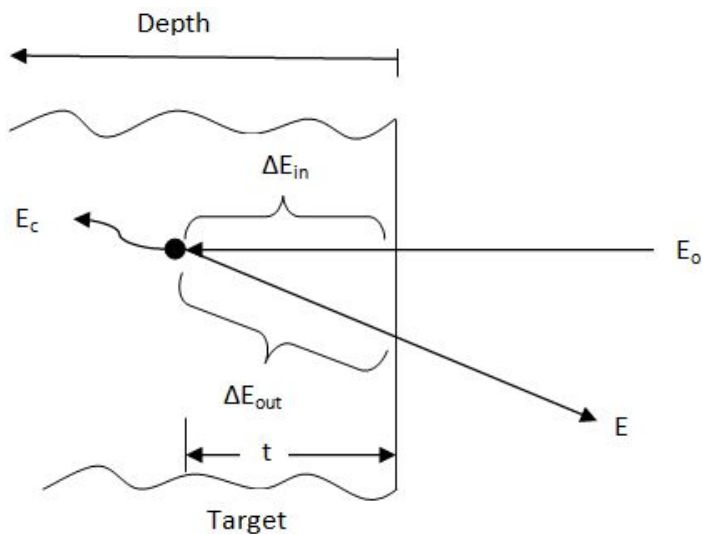
Schematic representation of the experimental setup for Rutherford backscattering analysis.

The probability of a scattering event can be described by the differential scattering cross section of a target atom for scattering an incoming particle through the angle θ into differential solid angle as follows,

$$\frac{d\sigma_R}{d\varphi} = \left(\frac{zZe^2}{2E_0 \sin^2 \theta} \right) \frac{\left[\cos \theta + \sqrt{1 - \left(\frac{M_1}{M_2} \sin \theta \right)^2} \right]^2}{\sqrt{1 - \left(\frac{M_1}{M_2} \sin \theta \right)^2}}$$

where $d\sigma_R$ is the effective differential cross section for the scattering of a particle. The above equation may look complicated but it conveys the message that the probability of scattering event can be expressed as a function of scattering cross section which is proportional to the zZ when a particle with charge ze approaches the target atom with charge Ze .

Helium ions not scattered at the surface lose energy as they traverse the solid. They lose energy due to interaction with electrons in the target. After collision the He particles lose further energy on their way out to the detector. We need to know two quantities to measure the energy loss, the distance Δt that the particles penetrate into the target and the energy loss ΔE in this distance [\[link\]](#). The rate of energy loss or stopping power is a critical component in backscattering experiments as it determines the depth profile in a given experiment.

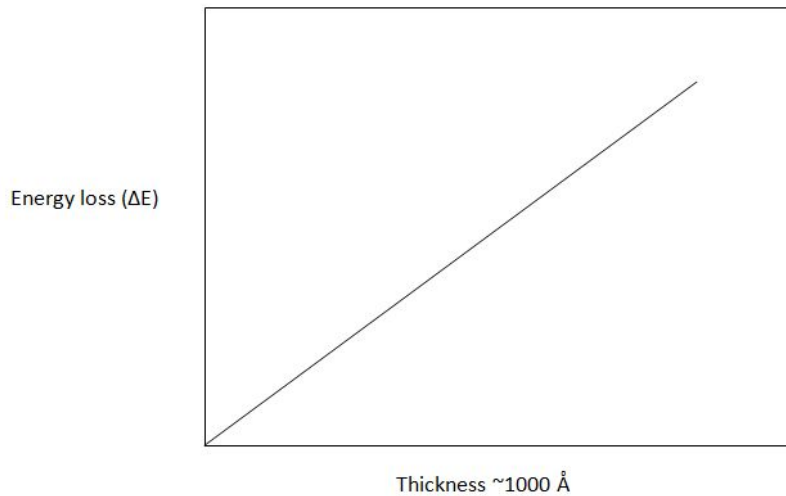


Components of energy loss for a ion beam that scatters from depth t . First, incident beam loses energy through interaction with electrons ΔE_{in} .

Then energy lost occurs due to scattering E_c . Finally outgoing beam loses energy for interaction with electrons ΔE_{out} . Adapted from L. C. Feldman and J. W. Mayer, *Fundamentals of Surface and Thin Film Analysis*, North Holland-Elsevier, New York (1986).

In thin film analysis, it is convenient to assume that total energy loss ΔE into depth t is only proportional to t for a given target. This assumption allows a simple derivation of energy

loss in backscattering as more complete analysis requires many numerical techniques. In constant dE/dx approximation, total energy loss becomes linearly related to depth t , [\[link\]](#).



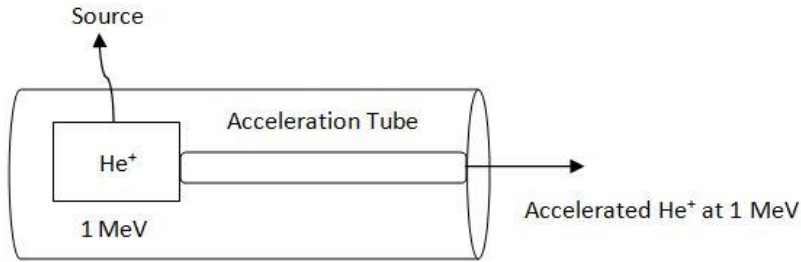
Variation of energy loss with the depth of the target in constant dE/dx approximation.

Experimental set-up

The apparatus for Rutherford backscattering analysis of thin solid surface typically consist of three components:

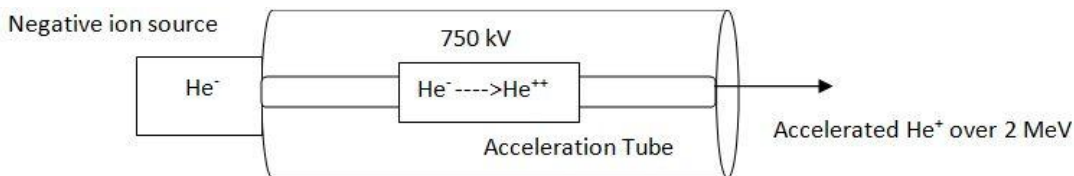
1. A source of helium ions.
2. An accelerator to energize the helium ions.
3. A detector to measure the energy of scattered ions.

There are two types of accelerator/ion source available. In single stage accelerator, the He^+ source is placed within an insulating gas-filled tank ([\[link\]](#)). It is difficult to install new ion source when it is exhausted in this type of accelerator. Moreover, it is also difficult to achieve particles with energy much more than 1 MeV since it is difficult to apply high voltages in this type of system.



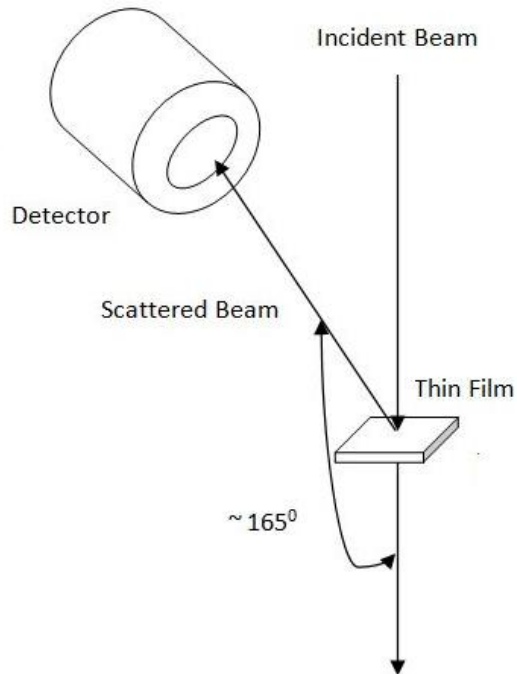
Schematic representation of a single stage accelerator.

Another variation is “tandem accelerator.” Here the ion source is at ground and produces negative ion. The positive terminal is located at the center of the acceleration tube ([\[link\]](#)). Initially the negative ion is accelerated from ground to terminal. At terminal two-electron stripping process converts the He^- to He^{++} . The positive ions are further accelerated toward ground due to columbic repulsion from positive terminal. This arrangement can achieve highly accelerated He^{++} ions ($\sim 2.25 \text{ MeV}$) with moderate voltage of 750 kV .



Schematic representation of a tandem accelerator.

Particles that are backscattered by surface atoms of the bombarded specimen are detected by a surface barrier detector. The surface barrier detector is a thin layer of p-type silicon on the n-type substrate resulting p-n junction. When the scattered ions exchange energy with the electrons on the surface of the detector upon reaching the detector, electrons get promoted from the valence band to the conduction band. Thus, each exchange of energy creates electron-hole pairs. The energy of scattered ions is detected by simply counting the number of electron-hole pairs. The energy resolution of the surface barrier detector in a standard RBS experiment is $12 - 20 \text{ keV}$. The surface barrier detector is generally set between 90° and 170° to the incident beam. Films are usually set normal to the incident beam. A simple layout is shown in [\[link\]](#).



Schematic representation general setup where the surface barrier detector is placed at angle of 165° to the extrapolated incident beam.

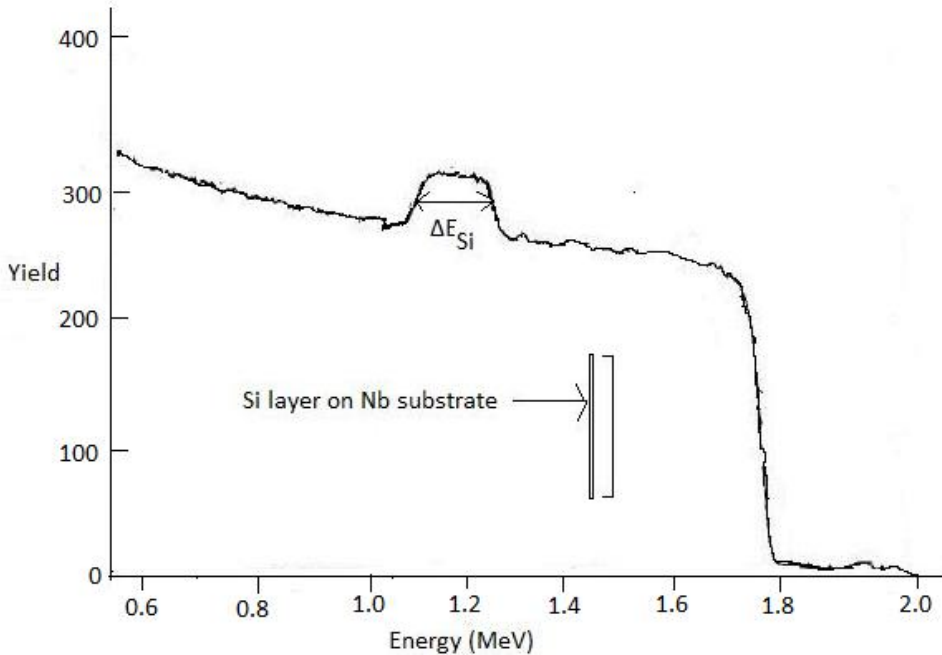
Depth profile analysis

As stated earlier, it is a good approximation in thin film analysis that the total energy loss ΔE is proportional to depth t . With this approximation, we can derive the relation between energy width ΔE of the signal from a film of thickness Δt as follows,

$$\Delta E = \Delta t(k \, dE/dx_{\text{in}} + 1/\cos\theta \, dE/dx_{\text{out}})$$

where θ = lab scattering angle.

It is worth noting that k is the kinematic factor defined in equation above and the subscripts “in” and “out” indicate the energies at which the rate of loss of energy or dE/dx is evaluated. As an example, we consider the backscattering spectrum, at scattering angle 170° , for 2 MeV He^{++} incidents on silicon layer deposited onto 2 mm thick niobium substrate [\[link\]](#).



The backscattering spectrum for 2.0 MeV He ions incident on a silicon thin film deposited onto a niobium substrate. Adapted from P. D. Stupik, M. M. Donovan, A. R. Barron, T. R. Jervis and M. Nastasi, *Thin Solid Films*, 1992, **207**, 138.

The energy loss rate of incoming He^{++} or dE/dx along inward path in elemental Si is $\approx 24.6 \text{ eV/\AA}$ at 2 MeV and is $\approx 26 \text{ eV/\AA}$ for the outgoing particle at 1.12 MeV (Since K of Si is 0.56 when the scattering angle is 170° , energy of the outgoing particle would be equal to 2×0.56 or 1.12 MeV). Again the value of ΔE_{Si} is $\approx 133.3 \text{ keV}$. Putting the values into above equation we get

$$\begin{aligned} \Delta t &\approx 133.3 \text{ keV} / (0.56 * 24.6 \text{ eV/\AA} + 1/\cos 170^\circ * 26 \text{ eV/\AA}) \\ &= 133.3 \text{ keV} / (13.77 \text{ eV/\AA} + 29.985 \text{ eV/\AA}) \\ &= 133.3 \text{ keV} / 40.17 \text{ eV/\AA} \\ &= 3318 \text{ \AA}. \end{aligned}$$

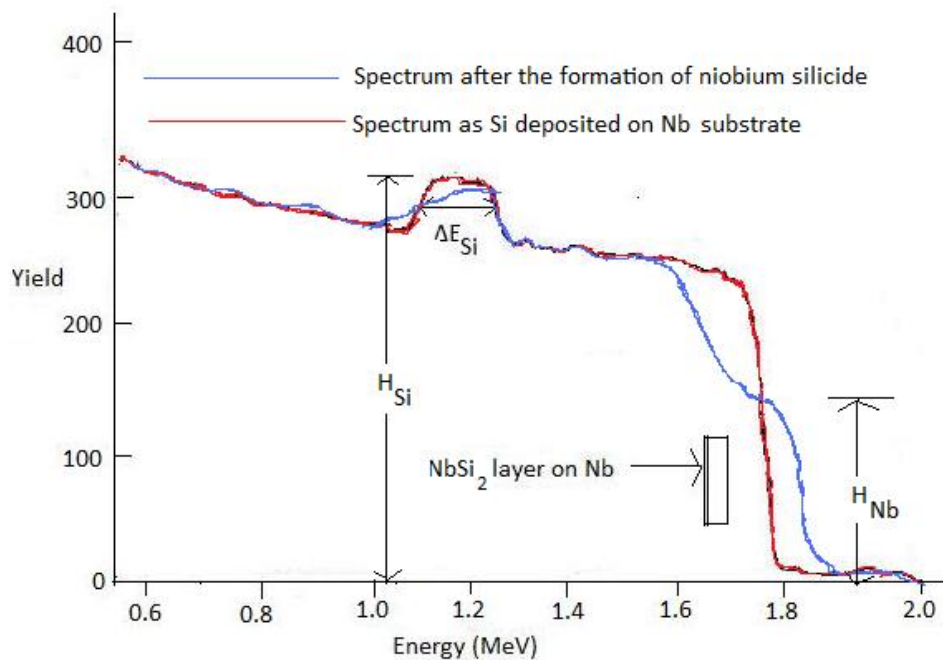
Hence a Si layer of ca. 3300 \AA thickness has been deposited on the niobium substrate. However we need to remember that the value of dE/dx is approximated in this calculation.

Quantitative Analysis

In addition to depth profile analysis, we can study the composition of an element quantitatively by backscattering spectroscopy. The basic equation for quantitative analysis is

$$Y = \sigma \cdot \Omega \cdot Q \cdot N\Delta t$$

Where Y is the yield of scattered ions from a thin layer of thickness Δt , Q is the number of incident ions and Ω is the detector solid angle, and $N\Delta t$ is the number of specimen atoms (atom/cm²). [\[link\]](#) shows the RBS spectrum for a sample of silicon deposited on a niobium substrate and subjected to laser mixing. The Nb has reacted with the silicon to form a NbSi₂ interphase layer. The Nb signal has broadened after the reaction as show in [\[link\]](#).



Backscattering spectra of Si diffused into Nb and Si as deposited on Nb substrate. Adapted from P. D. Stupik, M. M. Donovan, A. R. Barron, T. R. Jervis and M. Nastasi, *Thin Solid Films*, 1992, **207**, 138.

We can use ratio of the heights H_{Si}/H_{Nb} of the backscattering spectrum after formation of NbSi₂ to determine the composition of the silicide layer. The stoichiometric ratio of Nb and Si can be approximated as,

$$N_{Si}/N_{Nb} \approx [H_{Si} \cdot \sigma_{Si}]/[H_{Nb} \cdot \sigma_{Nb}]$$

Hence the concentration of Si and Nb can be determined if we can know the appropriate cross sections σ_{Si} and σ_{Nb} . However the yield in the backscattering spectra is better represented as the product of signal height and the energy width ΔE . Thus stoichiometric ratio can be better approximated as

$$N_{\text{Si}}/N_{\text{Nb}} \approx [H_{\text{Si}} * \Delta E_{\text{Si}} * \sigma_{\text{Si}}]/[H_{\text{Nb}} * \Delta E_{\text{Nb}} * \sigma_{\text{Nb}}]$$

Limitations

It is of interest to understand the limitations of the backscattering technique in terms of the comparison with other thin film analysis technique such as AES, XPS and SIMS ([\[link\]](#)). AES has better mass resolution, lateral resolution and depth resolution than RBS. But AES suffers from sputtering artifacts. Compared to RBS, SIMS has better sensitivity. RBS does not provide any chemical bonding information which we can get from XPS. Again, sputtering artifact problems are also associated in XPS. The strength of RBS lies in quantitative analysis. However, conventional RBS systems cannot analyze ultrathin films since the depth resolution is only about 10 nm using surface barrier detector.

Summary

Rutherford Backscattering analysis is a straightforward technique to determine the thickness and composition of thin films ($< 4000 \text{ \AA}$). Areas that have been lately explored are the use of backscattering technique in composition determination of new superconductor oxides; analysis of lattice mismatched epitaxial layers, and as a probe of thin film morphology and surface clustering.

Bibliography

- L. C. Feldman and J. W. Mayer, *Fundamentals of Surface and Thin Film Analysis*, North Holland-Elsevier, New York (1986).
- *Ion Spectroscopies for Surface Analysis*, Ed. A. W. Czanderna and D. M. Hercules, Plenum Press (New York), 1991.
- P. D. Stupik, M. M. Donovan, A. R Barron, T. R. Jervis, and M. Nastasi, *Thin Solid Films*, 1992, **207**, 138

The Application of VSI (Vertical Scanning Interferometry) to the Study of Crystal Surface Processes

Introduction

The processes which occur at the surfaces of crystals depend on many external and internal factors such as crystal structure and composition, conditions of a medium where the crystal surface exists and others. The appearance of a crystal surface is the result of complexity of interactions between the crystal surface and the environment. The mechanisms of surface processes such as dissolution or growth are studied by the physical chemistry of surfaces. There are a lot of computational techniques which allows us to predict the changing of surface morphology of different minerals which are influenced by different conditions such as temperature, pressure, pH and chemical composition of solution reacting with the surface. For example, Monte Carlo method is widely used to simulate the dissolution or growth of crystals. However, the theoretical models of surface processes need to be verified by natural observations. We can extract a lot of useful information about the surface processes through studying the changing of crystal surface structure under influence of environmental conditions. The changes in surface structure can be studied through the observation of crystal surface topography. The topography can be directly observed macroscopically or by using microscopic techniques. Microscopic observation allows us to study even very small changes and estimate the rate of processes by observing changing the crystal surface topography in time.

Much laboratory worked under the reconstruction of surface changes and interpretation of dissolution and precipitation kinetics of crystals. Invention of AFM made possible to monitor changes of surface structure during dissolution or growth. However, to detect and quantify the results of dissolution processes or growth it is necessary to determine surface area changes over a significantly larger field of view than AFM can provide. More recently, vertical scanning interferometry (VSI) has been developed as new tool to distinguish and trace the reactive parts of crystal surfaces. VSI and AFM are complementary techniques and practically well suited to detect surface changes.

VSI technique provides a method for quantification of surface topography at the angstrom to nanometer level. Time-dependent VSI measurements can be used to study the surface-normal retreat across crystal and other solid surfaces during dissolution process. Therefore, VSI can be used to directly and nondirectly measure mineral dissolution rates with high precision. Analogically, VSI can be used to study kinetics of crystal growth.

Physical principles of optical interferometry

Optical interferometry allows us to make extremely accurate measurements and has been used as a laboratory technique for almost a hundred years. Thomas Young observed interference of light and measured the wavelength of light in an experiment, performed around 1801. This experiment gave an evidence of Young's arguments for the wave model for light. The discovery of interference gave a basis to development of interferometry techniques widely successfully used as in microscopic investigations, as in astronomic investigations.

The physical principles of optical interferometry exploit the wave properties of light. Light can be thought as electromagnetic wave propagating through space. If we assume that we are dealing with a linearly polarized wave propagating in a vacuum in z direction, electric field E can be represented by a sinusoidal function of distance and time.

Equation:

$$E(x,y,z,t) = a \cos[2\pi(\nu t - z/\lambda)]$$

Where a is the amplitude of the light wave, ν is the frequency, and λ is its wavelength. The term within the square brackets is called the phase of the wave. Let's rewrite this equation in more compact form,

Equation:

$$E(x,y,z,t) = a \cos[\omega t - kz]$$

where $\omega = 2\pi\nu$ is the circular frequency, and $k = 2\pi/\lambda$ is the propagation constant. Let's also transform this second equation into a complex

exponential form,

Equation:

$$E(x,y,z,t) = \text{Re}\{a\exp(i\phi)\exp(i\omega t)\} = \text{Re}\{A\exp(i\omega t)\}$$

where $\phi = 2\pi z/\lambda$ and $A = \exp(-i\phi)$ is known as the complex amplitude. If n is a refractive index of a medium where the light propagates, the light wave traverses a distance d in such a medium. The equivalent optical path in this case is

Equation:

$$p = n \cdot d$$

When two light waves are superposed, the result intensity at any point depends on whether reinforce or cancel each other ([\[link\]](#)). This is well known phenomenon of interference. We will assume that two waves are propagating in the same direction and are polarized with their field vectors in the same plane. We will also assume that they have the same frequency. The complex amplitude at any point in the interference pattern is then the sum of the complex amplitudes of the two waves, so that we can write,

Equation:

$$A = A_1 + A_2$$

where $A_1 = a_1\exp(-i\phi_1)$ and $A_2 = a_2\exp(-i\phi_2)$ are the complex amplitudes of two waves. The resultant intensity is, therefore,

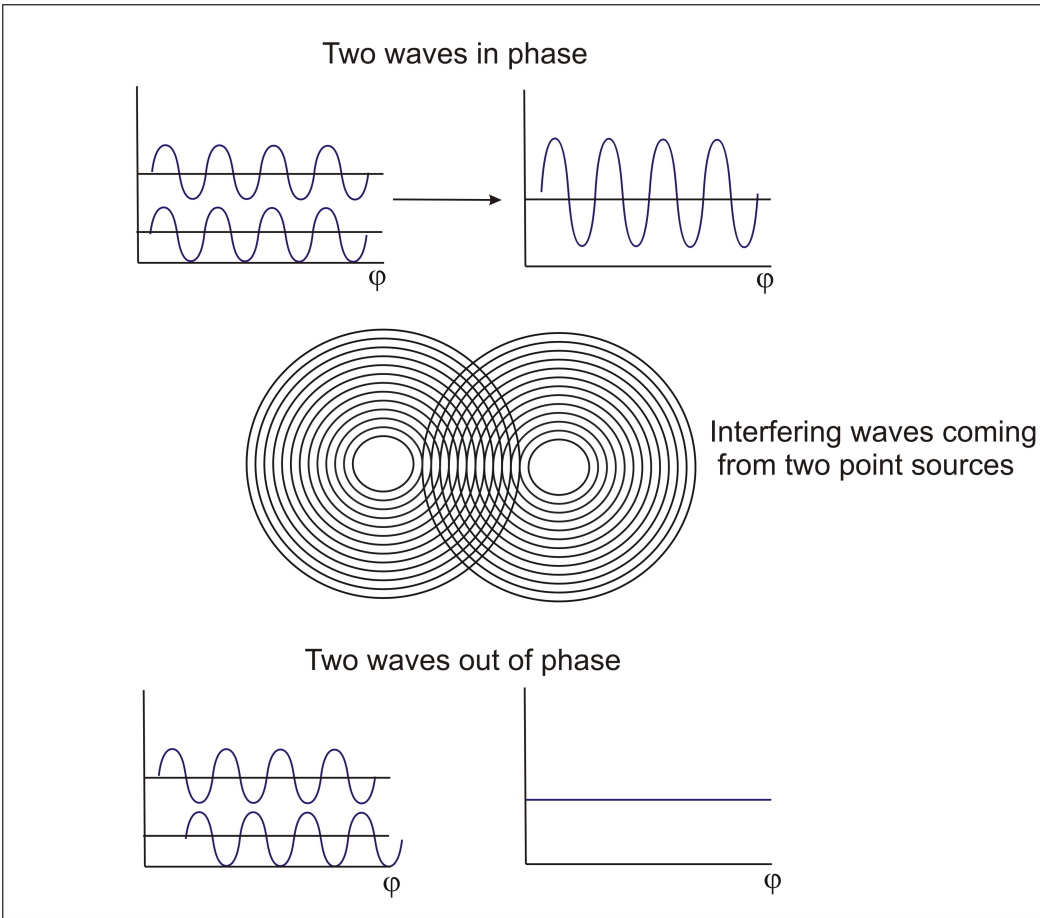
Equation:

$$I = |A|^2 = I_1 + I_2 + 2(I_1 I_2)^{1/2} \cos \Delta\phi$$

where I_1 and I_2 are the intensities of two waves acting separately, and $\Delta\phi = \phi_1 - \phi_2$ is the phase difference between them. If the two waves are derived from a common source, the phase difference corresponds to an optical path difference,

Equation:

$$\Delta p = (\lambda/2\pi)\Delta\phi$$



The scheme of interferometric wave interaction when two waves interact with each other, the amplitude of resulting wave will increase or decrease. The value of this amplitude depends on phase difference between two original waves.

If $\Delta\phi$, the phase difference between the beams, varies linearly across the field of view, the intensity varies cosinusoidally, giving rise to alternating light and dark bands or fringes ([\[link\]](#)). The intensity in an interference pattern has its maximum value

Equation:

$$I_{\max} = I_1 + I_2 + 2(I_1 I_2)^{1/2}$$

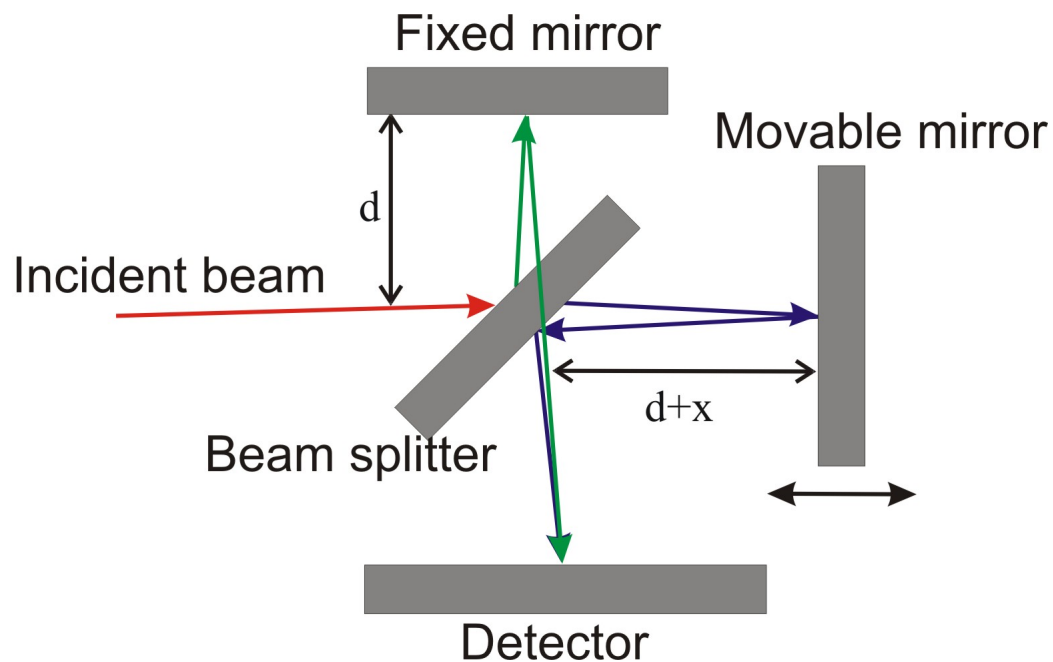
when $\Delta\phi = 2m\pi$, where m is an integer and its minimum value

Equation:

$$I_{\min} = I_1 + I_2 - 2(I_1 I_2)^{1/2}$$

when $\Delta\phi = (2m + 1)\pi$.

The principle of interferometry is widely used to develop many types of interferometric set ups. One of the earliest set ups is Michelson interferometry. The idea of this interferometry is quite simple: interference fringes are produced by splitting a beam of monochromatic light so that one beam strikes a fixed mirror and the other a movable mirror. An interference pattern results when the reflected beams are brought back together. The Michelson interferometric scheme is shown in [\[link\]](#).



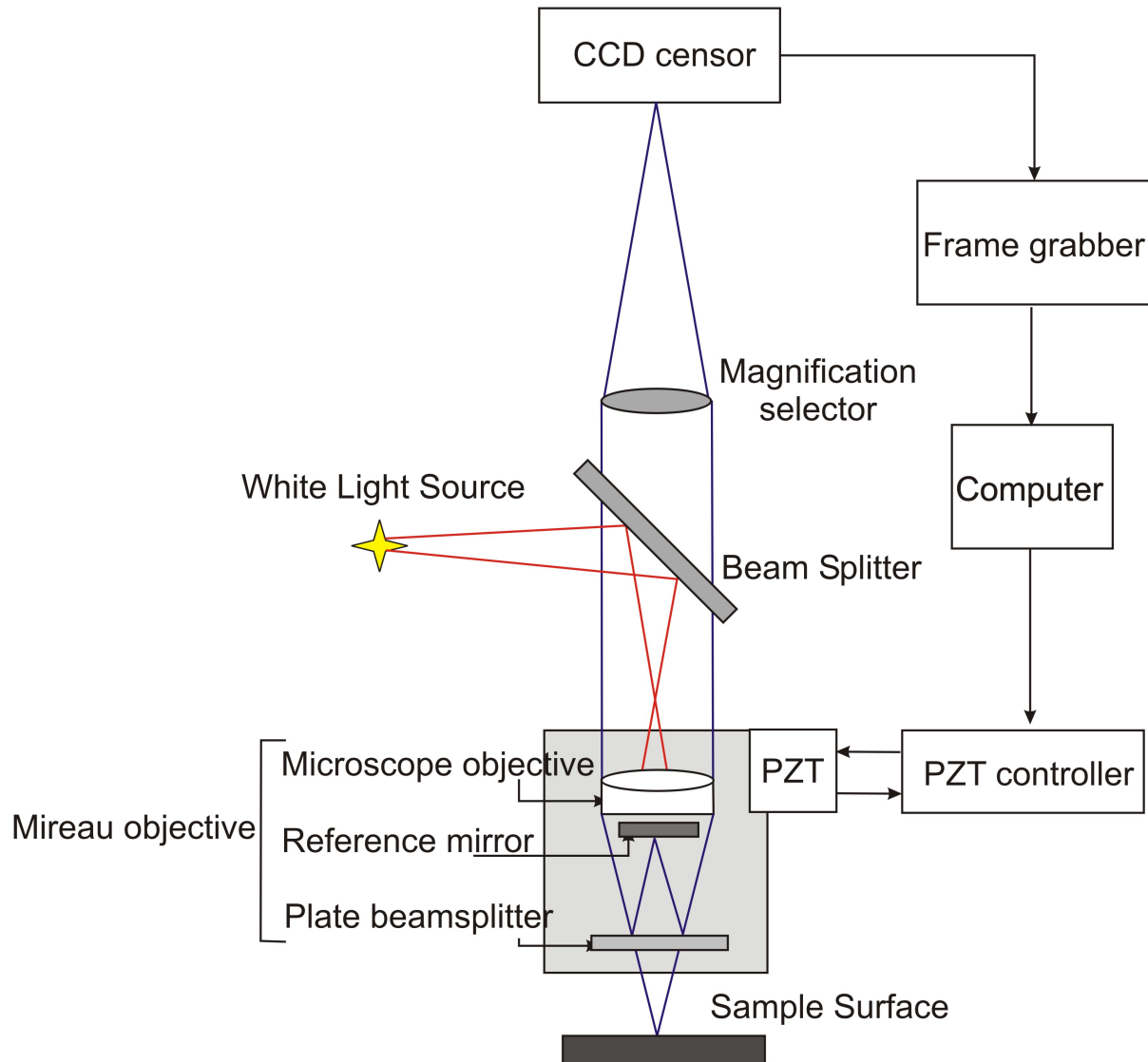
Schematic representation of a Michelson interferometry set-up.

The difference of path lengths between two beams is $2x$ because beams traverse the designated distances twice. The interference occurs when the path difference is equal to integer numbers of wavelengths,

Equation:

$$\Delta p = 2x = m\lambda, m = 0, \pm 1, \pm 2...$$

Modern interferometric systems are more complicated. Using special phase-measurement techniques they are capable to perform much more accurate height measurements than can be obtained just by directly looking at the interference fringes and measuring how they depart from being straight and equally spaced. Typically an interferometric system consists of a light source, beamsplitter, objective system, system of registration of signals and transformation into digital format and computer which processes data. Vertical scanning interferometry contains all these parts. [\[link\]](#) shows a configuration of VSI interferometric system.



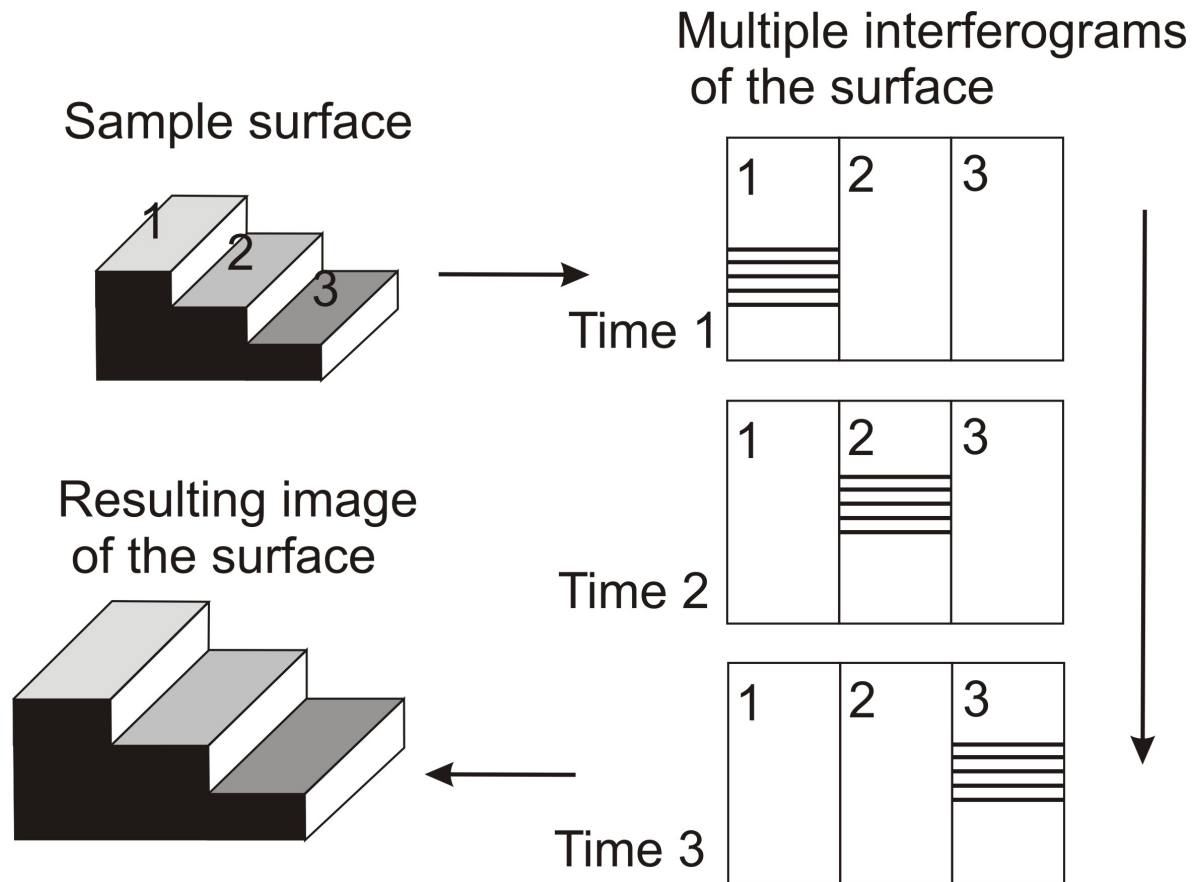
Schematic representation of the Vertical scanning interferometry (VSI) system.

Many of modern interferometric systems use Mirau objective in their constructions. Mirau objective is based on a Michelson interferometer. This objective consists of a lens, a reference mirror and a beamsplitter. The idea of getting interfering beams is simple: two beams (red lines) travel along the optical axis. Then they are reflected from the reference surface and the sample surface respectively (blue lines). After this these beams are

recombined to interfere with each other. An illumination or light source system is used to direct light onto a sample surface through a cube beam splitter and the Mireau objective. The sample surface within the field of view of the objective is uniformly illuminated by those beams with different incidence angles. Any point on the sample surface can reflect those incident beams in the form of divergent cone. Similarly, the point on the reference symmetrical with that on the sample surface also reflects those illuminated beams in the same form.

The Mireau objective directs the beams reflected of the reference and the sample surface onto a CCD (charge-coupled device) sensor through a tube lens. The CCD sensor is an analog shift register that enables the transportation of analog signals (electric charges) through successive stages (capacitors), controlled by a clock signal. The resulting interference fringe pattern is detected by CCD sensor and the corresponding signal is digitized by a frame grabber for further processing with a computer.

The distance between a minimum and a maximum of the interferogram produced by two beams reflected from the reference and sample surface is known. That is, exactly half the wavelength of the light source. Therefore, with a simple interferogram the vertical resolution of the technique would be also limited to $\lambda/2$. If we will use a laser light as a light source with a wavelength of 300 nm the resolution would be only 150 nm. This resolution is not good enough for a detailed near-atomic scale investigation of crystal surfaces. Fortunately, the vertical resolution of the technique can be improved significantly by moving either the reference or the sample by a fraction of the wavelength of the light. In this way, several interferograms are produced. Then they are all overlayed, and their phase shifts compared by the computer software [\[link\]](#). This method is widely known as phase shift interferometry (PSI).



Sketch illustrating phase-shift technology. The sample is continuously moved along the vertical axes in order to scan surface topography. All interferograms are automatically overlaid using computer software.

Most optical testing interferometers now use phase-shifting techniques not only because of high resolution but also because phase-shifting is a high accuracy rapid way of getting the interferogram information into the computer. Also usage of this technique makes the inherent noise in the data taking process very low. As the result in a good environment angstrom or sub-angstrom surface height measurements can be performed. As it was said above, in phase-shifting interferometry the phase difference between the interfering beams is changed at a constant rate as the detector is read out. Once the phase is determined across the interference field, the

corresponding height distribution on the sample surface can be determined. The phase distribution $\phi(x, y)$ is recorded by using the CCD camera.

Let's assign $A(x, y)$, $B(x, y)$, $C(x, y)$ and $D(x, y)$ to the resulting interference light intensities which are corresponded to phase-shifting steps of 0 , $\pi/2$, π and $3\pi/2$. These intensities can be obtained by moving the reference mirror through displacements of $\lambda/8$, $\lambda/4$ and $3\lambda/8$, respectively. The equations for the resulting intensities would be:

Equation:

$$A(x,y) = I_1(x,y) + I_2(x,y)\cos\alpha(x,y)$$

Equation:

$$B(x,y) = I_1(x,y) - I_2(x,y)\sin\alpha(x,y)$$

Equation:

$$C(x,y) = I_1(x,y) - I_2(x,y)\cos\alpha(x,y)$$

Equation:

$$D(x,y) = I_1(x,y) + I_2(x,y)\sin\alpha(x,y)$$

where $I_1(x,y)$ and $I_2(x,y)$ are two overlapping beams from two symmetric points on the test surface and the reference respectively. Solving equations [\[link\]](#)–[\[link\]](#), the phase map $\phi(x, y)$ of a sample surface will be given by the relation:

Equation:

$$\phi(x,y) = \frac{B(x,y) - D(x,y)}{A(x,y) - C(x,y)}$$

Once the phase is determined across the interference field pixel by pixel on a two-dimensional CCD array, the local height distribution/contour, $h(x, y)$, on the test surface is given by

Equation:

$$h(x,y) = \frac{\lambda}{4\pi} \phi(x,y)$$

Normally the resulted fringe can be in the form of a linear fringe pattern by adjusting the relative position between the reference mirror and sample surfaces. Hence any distorted interference fringe would indicate a local profile/contour of the test surface.

It is important to note that the Mireau objective is mounted on a capacitive closed-loop controlled PZT (piezoelectric actuator) as to enable phase shifting to be accurately implemented. The PZT is based on piezoelectric effect referred to the electric potential generated by applying pressure to piezoelectric material. This type of materials is used to convert electrical energy to mechanical energy and vice-versa. The precise motion that results when an electric potential is applied to a piezoelectric material has an importance for nanopositioning. Actuators using the piezo effect have been commercially available for 35 years and in that time have transformed the world of precision positioning and motion control.

Vertical scanning interferometer also has another name; white-light interferometry (WLI) because of using the white light as a source of light. With this type of source a separate fringe system is produced for each wavelength, and the resultant intensity at any point of examined surface is obtained by summing these individual patterns. Due to the broad bandwidth of the source the coherent length L of the source is short:

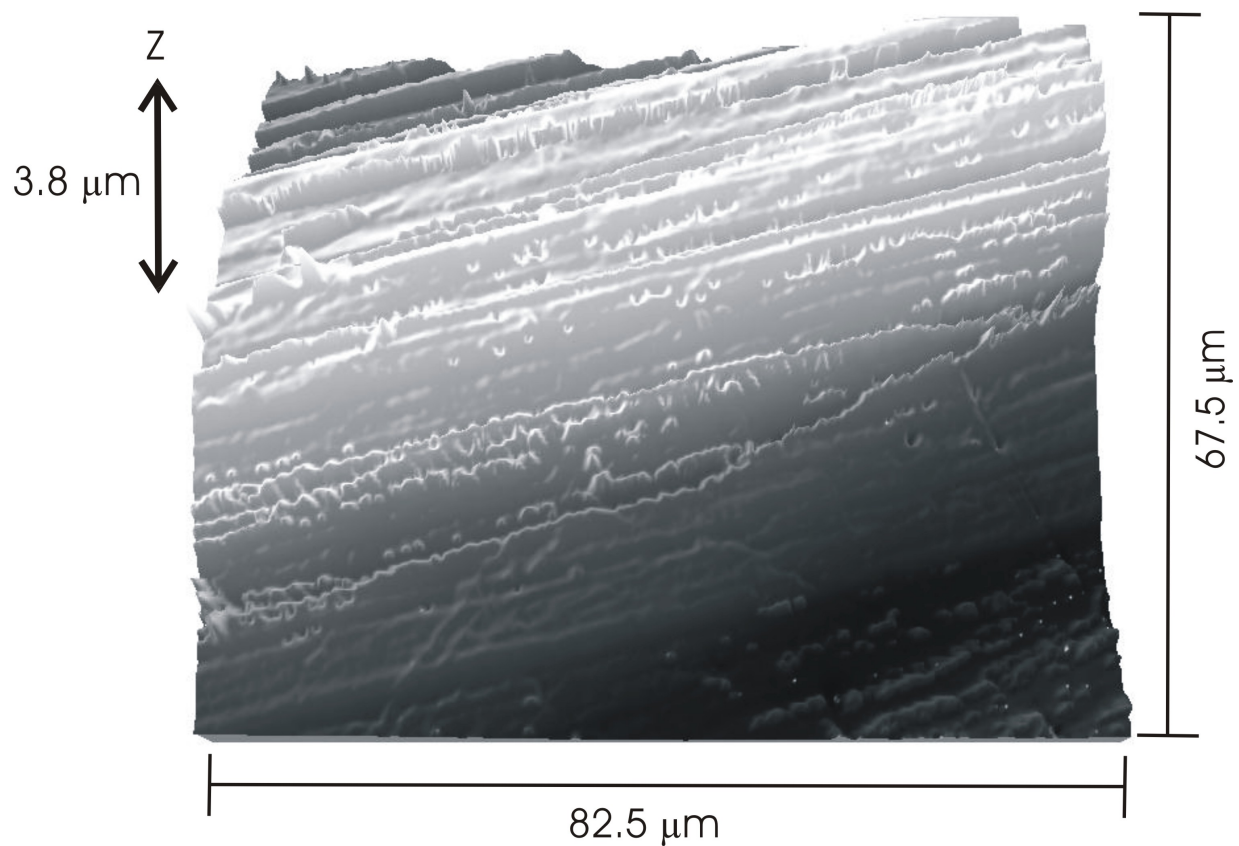
Equation:

$$L = \frac{\lambda^2}{n\Delta\lambda}$$

where λ is the center wavelength, n is the refractive index of the medium, $\Delta\lambda$ is the spectral width of the source. In this way good contrast fringes can be obtained only when the lengths of interfering beams pathways are closed to each other. If we will vary the length of a pathway of a beam reflected from sample, the height of a sample can be determined by looking at the position for which a fringe contrast is a maximum. In this case interference

pattern exist only over a very shallow depth of the surface. When we vary a pathway of sample-reflected beam we also move the sample in a vertical direction in order to get the phase at which maximum intensity of fringes will be achieved. This phase will be converted in height of a point at the sample surface.

The combination of phase shift technology with white-light source provides a very powerful tool to measure the topography of quite rough surfaces with the amplitude in heights about and the precision up to 1-2 nm. Through a developed software package for quantitatively evaluating the resulting interferogram, the proposed system can retrieve the surface profile and topography of the sample objects [\[link\]](#).



Example of muscovite surface topography, obtained by using VSI- 50x

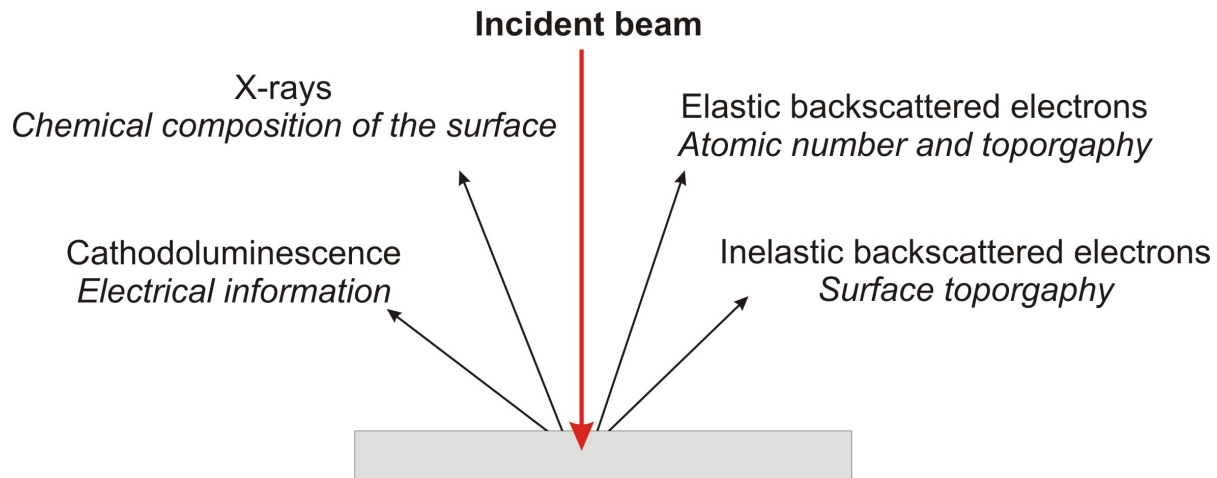
objective.

A comparison of common methods to determine surface topography: SEM, AFM and VSI

Except the interferometric methods described above, there are a several other microscopic techniques for studying crystal surface topography. The most common are scanning electron microscopy (SEM) and atomic force microscopy (AFM). All these techniques are used to obtain information about the surface structure. However they differ from each other by the physical principles on which they based.

Scanning electron microscopy

SEM allows us to obtain images of surface topography with the resolution much higher than the conventional light microscopes do. Also it is able to provide information about other surface characteristics such as chemical composition, electrical conductivity etc, see [\[link\]](#). All types of data are generated by the reflecting of accelerated electron beams from the sample surface. When electrons strike the sample surface, they lose their energy by repeated random scattering and adsorption within an outer layer into the depth varying from 100 nm to 5 microns.



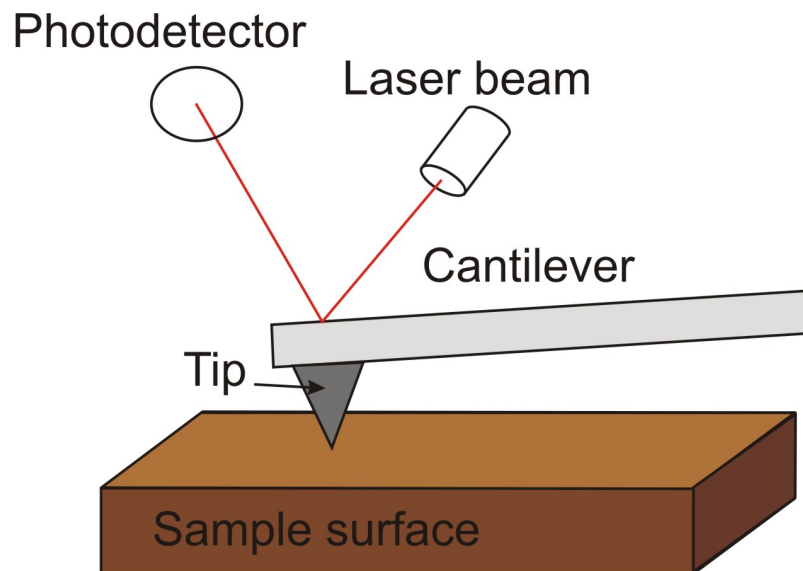
Scheme of electron beam-sample interaction at SEM analysis

The thickness of this outer layer also known as interactive layer depends on energy of electrons in the beam, composition and density of a sample. Result of the interaction between electron beam and the surface provides several types of signals. The main type is secondary or inelastic scattered electrons. They are produced as a result of interaction between the beam of electrons and weakly bound electrons in the conduction band of the sample. Secondary electrons are ejected from the k-orbitals of atoms within the surface layer of thickness about a few nanometers. This is because secondary electrons are low energy electrons (<50 eV), so only those formed within the first few nanometers of the sample surface have enough energy to escape and be detected. Secondary backscattered electrons provide the most common signal to investigate surface topography with lateral resolution up to 0.4 - 0.7 nm.

High energy beam electrons are elastic scattered back from the surface. This type of signal gives information about chemical composition of the surface because the energy of backscattered electrons depends on the weight of atoms within the interaction layer. Also this type of electrons can form secondary electrons and escape from the surface or travel further into the sample than the secondary. The SEM image formed is the result of the intensity of the secondary electron emission from the sample at each x,y data point during the scanning of the surface.

Atomic force microscopy

AFM is a very popular tool to study surface dissolution. AFM set up consists of scanning a sharp tip on the end of a flexible cantilever which moves across a sample surface. The tips typically have an end radius of 2 to 20 nm, depending on tip type. When the tip touches the surface the forces of these interactions leads to deflection of a cantilever. The interaction between tip and sample surface involve mechanical contact forces, van der Waals forces, capillary forces, chemical bonding, electrostatic forces, magnetic forces etc. The deflection of a cantilever is usually measured by reflecting a laser beam off the back of the cantilever into a split photodiode detector. A schematic drawing of AFM can be seen in [\[link\]](#). The two most commonly used modes of operation are contact mode AFM and tapping mode AFM, which are conducted in air or liquid environments.



Schematic drawing of an AFM apparatus.

Working under the contact mode AFM scans the sample while monitoring the change in cantilever deflection with the split photodiode detector. Loop maintains a constant cantilever reflection by vertically moving the scanner

to get a constant signal. The distance which the scanner goes by moving vertically at each x,y data point is stored by the computer to form the topographic image of the sample surface. Working under the tapping mode AFM oscillates the cantilever at its resonance frequency (typically ~300 kHz) and lightly “taps” the tip on the surface during scanning. The electrostatic forces increase when tip gets close to the sample surface, therefore the amplitude of the oscillation decreases. The laser deflection method is used to detect the amplitude of cantilever oscillation. Similar to the contact mode, feedback loop maintains a constant oscillation amplitude by moving the scanner vertically at every x,y data point. Recording this movement forms the topographical image. The advantage of tapping mode over contact mode is that it eliminates the lateral, shear forces present in contact mode. This enables tapping mode to image soft, fragile, and adhesive surfaces without damaging them while work under contact mode allows the damage to occur.

Comparison of techniques

All techniques described above are widely used in studying of surface nano- and micromorphology. However, each method has its own limitations and the proper choice of analytical technique depends on features of analyzed surface and primary goals of research.

All these techniques are capable to obtain an image of a sample surface with quite good resolution. The lateral resolution of VSI is much less, then for other techniques: 150 nm for VSI and 0.5 nm for AFM and SEM. Vertical resolution of AFM (0.5 Å) is better then for VSI (1 - 2 nm), however VSI is capable to measure a high vertical range of heights (1 mm) which makes possible to study even very rough surfaces. On the contrary, AFM allows us to measure only quite smooth surfaces because of its relatively small vertical scan range (7 µm). SEM has less resolution, than AFM because it requires coating of a conductive material with the thickness within several nm.

The significant advantage of VSI is that it can provide a large field of view ($845 \times 630 \mu\text{m}$ for 10x objective) of tested surfaces. Recent studies of

surface roughness characteristics showed that the surface roughness parameters increase with the increasing field of view until a critical size of 250,000 μm is reached. This value is larger than the maximum field of view produced by AFM ($100 \times 100 \mu\text{m}$) but can be easily obtained by VSI. SEM is also capable to produce images with large field of view. However, SEM is able to provide only 2D images from one scan while AFM and VSI let us to obtain 3D images. It makes quantitative analysis of surface topography more complicated, for example, topography of membranes is studied by cross section and top view images.

	VSI	AFM	SEM
Lateral resolution	0.5-1.2 μm	0.5 nm	0.5-1 nm
Vertical resolution	2 nm	0.5 Å	Only 2D images
Field of view	845 \times 630 μm (10x objective)	100 \times 100 μm	1-2 mm
Vertical range of scan	1 mm	10 μm	-
Preparation of a sample	-	-	Required coating of a conducted material
Required environment	Air	Air, liquid	Vacuum

A comparison of VSI sample and resolution with AFM and SEM.

The experimental studying of surface processes using microscopic techniques

The limitations of each technique described above are critically important to choose appropriate technique for studying surface processes. Let's explore application of these techniques to study dissolution of crystals.

When crystalline matter dissolves the changes of the crystal surface topography can be observed by using microscopic techniques. If we will apply an unreactive mask (silicon for example) on crystal surface and place a crystalline sample into the experiment reactor then we get two types of surfaces: dissolving and remaining the same or unreacted. After some period of time the crystal surface starts to dissolve and change its z-level. In order to study these changes *ex situ* we can pull out a sample from the reaction cell then remove a mask and measure the average height difference $\Delta\bar{h}$ between the unreacted and dissolved areas. The average heights of dissolved and unreacted areas are obtained through digital processing of data obtained by microscopes. The velocity of normal surface retreat v_{SNR} during the time interval Δt is defined as

$$v_{\text{SNR}} = \frac{\Delta\bar{h}}{\Delta t}$$

Dividing this velocity by the molar volume $V(\text{cm}^3/\text{mol})$ gives a global dissolution rate in the familiar units of moles per unit area per unit time:

Equation:

$$R = \frac{v_{\text{SNR}}}{V}$$

This method allows us to obtain experimental values of dissolution rates just by precise measuring of average surface heights. Moreover, using this method we can measure local dissolution rates at etch pits by monitoring changes in the volume and density of etch pits across the surface over time. VSI technique is capable to perform these measurements because of large vertical range of scanning. In order to get precise values of rates which are

not depend on observing place of crystal surface we need to measure enough large areas. VSI technique provides data from areas which are large enough to study surfaces with heterogeneous dissolution dynamics and obtain average dissolution rates. Therefore, VSI makes possible to measure rates of normal surface retreat during the dissolution and observe formation, growth and distribution of etch pits on the surface.

However, if the mechanism of dissolution is controlled by dynamics of atomic steps and kink sites within a smooth atomic surface area, the observation of the dissolution process need to use a more precise technique. AFM is capable to provide information about changes in step morphology *in situ* when the dissolution occurs. For example, immediate response of the dissolved surface to the changing of environmental conditions (concentrations of ions in the solution, pH etc.) can be studied by using AFM.

SEM is also used to examine micro and nanotexture of solid surfaces and study dissolution processes. This method allows us to observe large areas of crystal surface with high resolution which makes possible to measure a high variety of surfaces. The significant disadvantage of this method is the requirement to cover examine sample by conductive substance which limits the resolution of SEM. The other disadvantage of SEM is that the analysis is conducted in vacuum. Recent technique, environmental SEM or ESEM overcomes these requirements and makes possible even examine liquids and biological materials. The third disadvantage of this technique is that it produces only 2D images. This creates some difficulties to measure Δh within the dissolving area. One of advantages of this technique is that it is able to measure not only surface topography but also chemical composition and other surface characteristics of the surface. This fact is used to monitor changing in chemical composition during the dissolution.

Bibliography

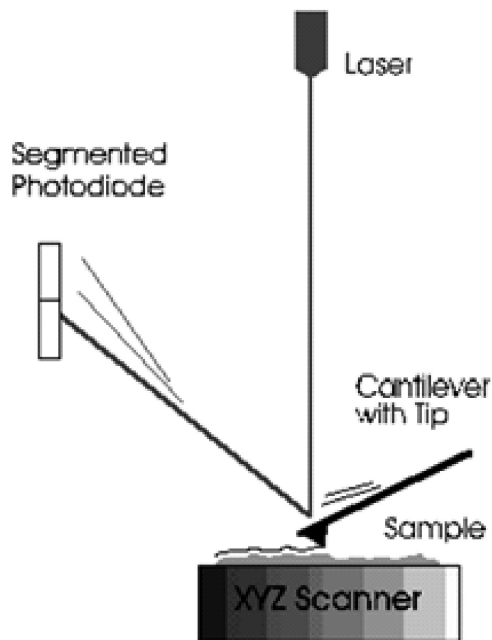
- A. C. Lasaga, *Kinetic Theory in the Earth Sciences*. Princeton Univ. Press, Princeton, NJ (1998).
- A. Luttge, E. V. Bolton, and A. C. Lasaga A.C., *Am. J. Sci.*, 1999, **299**, 652.

- D. Kaczmarek, *Vacuum*, 2001, **62**, 303.
- P. Hariharan. *Optical interferometry*, Second edition, Academic press (2003) ISBN 0-12-311630-9.
- A. Luttge and P. G. Conrad, *Appl. Environ. Microbiol.*, 2004, **70**, 1627.
- A. C. Lasaga and A. Luttge, *American Mineralogist*, 2004, **89**, 527.
- K. J. Davis and A. Luttge, *Am. J. Sci.*, 2005, **305**, 727.
- S. H. Wang and Tay, *Meas. Sci. Technol.*, 2006, **17**, 617.
- A. Luttge and R. S. Arvidson, in *Kinetics of water-rock interaction*, Ed. S. Brantley, J. Kubicki, and A. White, Springer (2007).
- L. Zhang and A. Luttge, *American Mineralogist*, 2007, **92**, 1316.
- C. Fischer A. and Luttge, *Am. J. Sci.*, 2007, **307**, 955.
- Y. Wyart, G. Georges, C. Deumie, C. Amra, and P. Moulina, *J. Membrane Sci.*, 2008, **315**, 82.
- T. C. Vaimakis, E. D. Economou, and C. C. Trapalis, *J. Therm. Anal. Cal.*, 2008, **92**, 783.

Atomic Force Microscopy

Introduction

Atomic force microscopy (AFM) is a high-resolution form of scanning probe microscopy, also known as scanning force microscopy (SFM). The instrument uses a cantilever with a sharp tip at the end to scan over the sample surface ([\[link\]](#)). As the probe scans over the sample surface, attractive or repulsive forces between the tip and sample, usually in the form of van der Waal forces but also can be a number of others such as electrostatic and hydrophobic/hydrophilic, cause a deflection of the cantilever. The deflection is measured by a laser ([\[link\]](#)) which is reflected off the cantilever into photodiodes. As one of the photodiodes collects more light, it creates an output signal that is processed and provides information about the vertical bending of the cantilever. This data is then sent to a scanner that controls the height of the probe as it moves across the surface. The variance in height applied by the scanner can then be used to produce a three-dimensional topographical representation of the sample.



Simple schematic of

atomic force microscope
(AFM) apparatus.
Adapted from H. G.
Hansma, Department of
Physics, University of
California, Santa Barbara.

Modes of operation

Contact mode

The contact mode method utilizes a constant force for tip-sample interactions by maintaining a constant tip deflection ([\[link\]](#)). The tip communicates the nature of the interactions that the probe is having at the surface via feedback loops and the scanner moves the entire probe in order to maintain the original deflection of the cantilever. The constant force is calculated and maintained by using Hooke's Law, [\[link\]](#). This equation relates the force (F), spring constant (k), and cantilever deflection (x). Force constants typically range from 0.01 to 1.0 N/m. Contact mode usually has the fastest scanning times but can deform the sample surface. It is also only the only mode that can attain "atomic resolution."

Equation:

$$F = -kx$$



Schematic diagram of

probe and surface
interaction in contact
mode.

Tapping mode

In the tapping mode the cantilever is externally oscillated at its fundamental resonance frequency ([link](#)). A piezoelectric on top of the cantilever is used to adjust the amplitude of oscillation as the probe scans across the surface. The deviations in the oscillation frequency or amplitude due to interactions between the probe and surface are measured, and provide information about the surface or types of material present in the sample. This method is gentler than contact AFM since the tip is not dragged across the surface, but it does require longer scanning times. It also tends to provide higher lateral resolution than contact AFM.

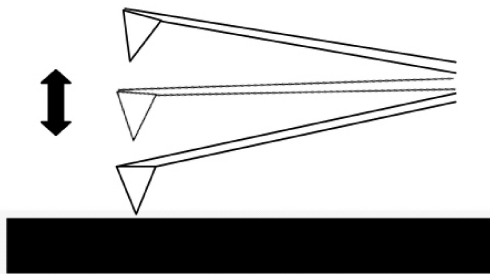


Diagram of probe and
surface interaction in
tapping mode.

Noncontact mode

For noncontact mode the cantilever is oscillated just above its resonance frequency and this frequency is decreased as the tip approaches the surface and experiences the forces associated with the material ([\[link\]](#)). The average tip-to-sample distance is measured as the oscillation frequency or amplitude is kept constant, which then can be used to image the surface. This method exerts very little force on the sample, which extends the lifetime of the tip. However, it usually does not provide very good resolution unless placed under a strong vacuum.

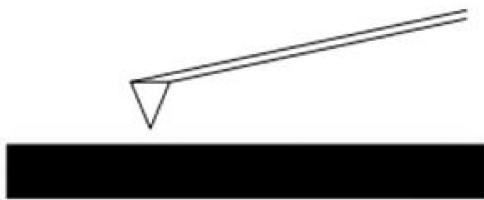


Diagram of probe and surface interaction in noncontact mode.

Experimental limitations

A common problem seen in AFM images is the presence of artifacts which are distortions of the actual topography, usually either due to issues with the probe, scanner, or image processing. The AFM scans slowly which makes it more susceptible to external temperature fluctuations leading to thermal drift. This leads to artifacts and inaccurate distances between topographical features.

It is also important to consider that the tip is not perfectly sharp and therefore may not provide the best aspect ratio, which leads to a convolution of the true topography. This leads to features appearing too large or too small since the width of the probe cannot precisely move around the particles and holes on the surface. It is for this reason that tips

with smaller radii of curvature provide better resolution in imaging. The tip can also produce false images and poorly contrasted images if it is blunt or broken.

The movement of particles on the surface due to the movement of the cantilever can cause noise, which forms streaks or bands in the image. Artifacts can also be made by the tip being of inadequate proportions compared to the surface being scanned. It is for this reason that it is important to use the ideal probe for the particular application.

Sample size and preparation

The sample size varies with the instrument but a typical size is 8 mm by 8 mm with a typical height of 1 mm. Solid samples present a problem for AFM since the tip can shift the material as it scans the surface. Solutions or dispersions are best for applying as uniform of a layer of material as possible in order to get the most accurate value of particles' heights. This is usually done by spin-coating the solution onto freshly cleaved mica which allows the particles to stick to the surface once it has dried.

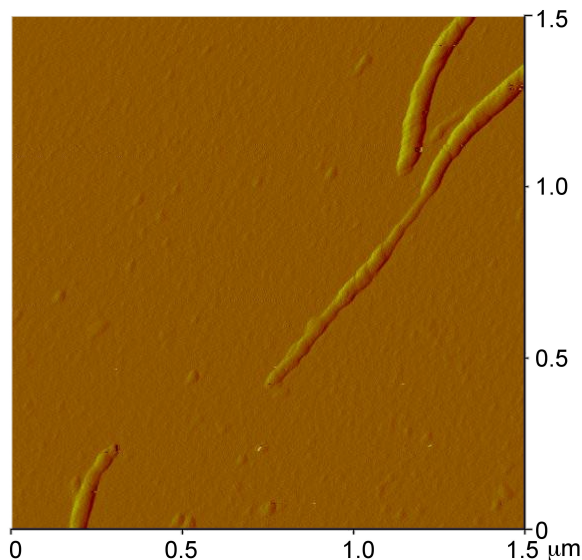
Applications of AFM

AFM is particularly versatile in its applications since it can be used in ambient temperatures and many different environments. It can be used in many different areas to analyze different kinds of samples such as semiconductors, polymers, nanoparticles, biotechnology, and cells amongst others. The most common application of AFM is for morphological studies in order to attain an understanding of the topography of the sample. Since it is common for the material to be in solution, AFM can also give the user an idea of the ability of the material to be dispersed as well as the homogeneity of the particles within that dispersion. It also can provide a lot of information about the particles being studied such as particle size, surface area, electrical properties, and chemical composition. Certain tips are capable of determining the principal mechanical, magnetic, and electrical properties of the material. For example, in magnetic force microscopy (MFM) the probe has a magnetic coating that senses magnetic, electrostatic, and atomic interactions with the surface. This type of scanning can be

performed in static or dynamic mode and depicts the magnetic structure of the surface.

AFM of carbon nanotubes

Atomic force microscopy is usually used to study the topographical morphology of these materials. By measuring the thickness of the material it is possible to determine if bundling occurred and to what degree. Other dimensions of the sample can also be measured such as the length and width of the tubes or bundles. It is also possible to detect impurities, functional groups ([\[link\]](#)), or remaining catalyst by studying the images. Various methods of producing nanotubes have been found and each demonstrates a slightly different profile of homogeneity and purity. These impurities can be carbon coated metal, amorphous carbon, or other allotropes of carbon such as fullerenes and graphite. These facts can be utilized to compare the purity and homogeneity of the samples made from different processes, as well as monitor these characteristics as different steps or reactions are performed on the material. The distance between the tip and the surface has proven itself to be an important parameter in noncontact mode AFM and has shown that if the tip is moved past the threshold distance, approximately 30 μm , it can move or damage the nanotubes. If this occurs, a useful characterization cannot be performed due to these distortions of the image.



AFM image of a polyethyleneimine-functionalized single walled carbon nanotube (PEI-SWNT) showing the presence of PEI “globules” on the SWNT. Adapted from E. P. Dillon, C. A. Crouse, and A. R. Barron, *ACS Nano*, 2008, 2, 156.

AFM of fullerenes

Atomic force microscopy is best applied to aggregates of fullerenes rather than individual ones. While the AFM can accurately perform height analysis of individual fullerene molecules, it has poor lateral resolution and it is difficult to accurately depict the width of an individual molecule. Another common issue that arises with contact AFM and fullerene deposited films is that the tip shifts clusters of fullerenes which can lead to discontinuities in sample images.

Bibliography

- R. Anderson and A. R. Barron, *J. Am. Chem. Soc.*, 2005, **127**, 10458.
- M. Bellucci, G. Gaggiotti, M. Marchetti, F. Micciulla, R. Mucciato, and M. Regi, *J. Physics: Conference Series*, 2007, **61**, 99.
- I. I. Bobrinetskii, V. N. Kukin, V. K. Nevolin, and M. M. Simunin. *Semiconductor*, 2008, **42**, 1496.
- S. H. Cohen and M. L. Lightbody. *Atomic Force Microscopy/Scanning Tunneling Microscopy 2*. Plenum, New York (1997).
- E. P. Dillon, C. A. Crouse, and A. R. Barron, *ACS Nano*, 2008, **2**, 156.
- C. Gu, C. Ray, S. Guo, and B. B. Akhremitchev, *J. Phys. Chem.*, 2007, **111**, 12898.
- G. Kaupp, *Atomic Force Microscopy, Scanning Nearfield Optical Microscopy and Nanoscratching: Application to Rough and Natural Surfaces*. Springer-Verlag, Berlin (2006).
- S. Morita, R. Wiesendanger, E. Meyer, and F. J. Giessibl. *Noncontact Atomic Force Microscopy*. Springer, Berlin (2002).

Introduction to Nanoparticle Synthesis

The fabrication of nanomaterials with strict control over size, shape, and crystalline structure has inspired the application of nanochemistry to numerous fields including catalysis, medicine, and electronics. The use of nanomaterials in such applications also requires the development of methods for nanoparticle assembly or dispersion in various media. A majority of studies have been aimed at dispersion in aqueous media aimed at their use in medical applications and studies of environmental effects, however, the principles of nanoparticle fabrication and functionalization of nanoparticles transcends their eventual application. Herein, we review the most general routes to nanoparticles of the key types that may have particular application within the oil and gas industry for sensor, composite, or device applications.

Synthesis methods for nanoparticles are typically grouped into two categories: “top-down” and “bottom-up”. The first involves division of a massive solid into smaller portions. This approach may involve milling or attrition, chemical methods, and volatilization of a solid followed by condensation of the volatilized components. The second, “bottom-up”, method of nanoparticle fabrication involves condensation of atoms or molecular entities in a gas phase or in solution. The latter approach is far more popular in the synthesis of nanoparticles.

Dispersions of nanoparticles are intrinsically thermodynamically metastable, primarily due to their very high surface area, which represents a positive contribution to the free enthalpy of the system. If the activation energies are not sufficiently high, evolution of the nanoparticle dispersion occurs causing an increase in nanoparticle size as typified by an Ostwald ripening process. Thus, highly dispersed nanoparticles are only kinetically stabilized and cannot be prepared under conditions that exceed some threshold, meaning that so-called “soft-chemical” or “*chemie duce*” methods are preferred. In addition, the use of surface stabilization is employed in many nanomaterials to hinder sintering, recrystallization and aggregation.

Bibliography

- J. Gopalakrishnan, *Chem. Mater.*, 1995, 7, 1265.

Synthesis of Semiconductor Nanoparticles

The most studied non-oxide semiconductors are cadmium chalcogenides (CdE, with E = sulfide, selenide and telluride). CdE nanocrystals were probably the first material used to demonstrate quantum size effects corresponding to a change in the electronic structure with size, i.e., the increase of the band gap energy with the decrease in size of particles ([link](#)). These semiconductor nanocrystals are commonly synthesized by thermal decomposition of an organometallic precursor dissolved in an anhydrous solvent containing the source of chalcogenide and a stabilizing material (polymer or capping ligand). Stabilizing molecules bound to the surface of particles control their growth and prevent particle aggregation.

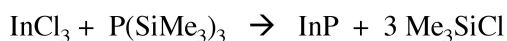


Picture of cadmium selenide (CdSe) quantum dots, dissolved in toluene, fluorescing brightly, as they are exposed to an ultraviolet lamp, in three noticeable different colors (blue ~481 nm, green ~520 nm, and orange ~612 nm) due to the quantum dots' bandgap (and thus the wavelength of emitted light) depends strongly on the particle size; the smaller the dot,

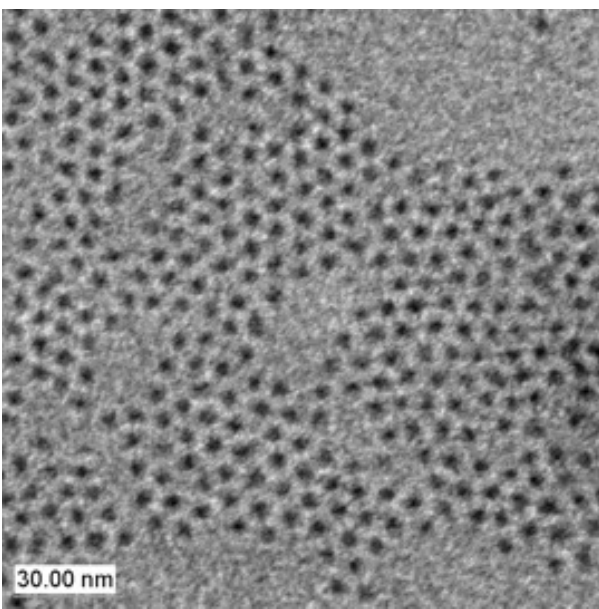
the shorter the emitted wavelength of light. The "blue" quantum dots have the smallest particle size, the "green" dots are slightly larger, and the "orange" dots are the largest.

Although cadmium chalcogenides are the most studied semiconducting nanoparticles, the methodology for the formation of semiconducting nanoparticles was first demonstrated independently for InP and GaAs, e.g., [\[link\]](#). This method has been adapted for a range of semiconductor nanoparticles.

Equation:



In the case of CdSe, dimethylcadmium $\text{Cd}(\text{CH}_3)_2$ is used as a cadmium source and bis(trimethylsilyl)sulfide, $(\text{Me}_3\text{Si})_2\text{S}$, trioctylphosphine selenide or telluride (TOPSe, TOPTe) serve as sources of selenide in trioctylphosphine oxide (TOPO) used as solvent and capping molecule. The mixture is heated at 230-260 °C over a few hours while modulating the temperature in response to changes in the size distribution as estimated from the absorption spectra of aliquots removed at regular intervals. These particles, capped with TOP/TOPO molecules, are non-aggregated ([\[link\]](#)) and easily dispersible in organic solvents forming optically clear dispersions. When similar syntheses are performed in the presence of surfactant, strongly anisotropic nanoparticles are obtained, e.g., rod-shaped CdSe nanoparticles can be obtained.



TEM image of CdSe nanoparticles.

Because $\text{Cd}(\text{CH}_3)_2$ is extremely toxic, pyrophoric and explosive at elevated temperature, other Cd sources have been used. CdO appears to be an interesting precursor. CdO powder dissolves in TOPO and HPA or TDPA (tetradecylphosphonic acid) at about 300 °C giving a colorless homogeneous solution. By introducing selenium or tellurium dissolved in TOP, nanocrystals grow to the desired size.

Nanorods of CdSe or CdTe can also be produced by using a greater initial concentration of cadmium as compared to reactions for nanoparticles. This approach has been successfully applied for synthesis of numerous other metal chalcogenides including ZnS, ZnSe, and $\text{Zn}_{1-x}\text{Cd}_x\text{S}$. Similar procedures enable the formation of MnS, PdS, NiS, Cu_2S nanoparticles, nano rods, and nano disks.

Bibliography

- C. R. Berry, *Phys. Rev.*, 1967, **161**, 848.

- M. D. Healy, P. E. Laibinis, P. D. Stupik, and A. R. Barron, *J. Chem. Soc., Chem. Commun.*, 1989, 359.
- L. Manna, E. C. Scher, and A. P. Alivisatos, *J. Am. Chem. Soc.*, 2000, **122**, 12700.
- C. B. Murray, D. J. Norris, and M. G. Bawendi, *J. Am. Chem. Soc.*, 1993, **115**, 8706.
- Z. A. Peng and X. Peng, *J. Am. Chem. Soc.*, 2002, **12**, 3343.
- R. L. Wells, C. G. Pitt, A. T. McPhail, A. P. Purdy, S. R. B. Shafieezad, and Hallock *Chem. Mater.*, 1989, **1**, 4.
- X. Zong, Y. Feng, W. Knoll, and H. Man, *J. Am. Chem. Soc.*, 2003, **125**, 13559.

Optical Properties of Group 12-16 (II-VI) Semiconductor Nanoparticles

What are Group 12-16 semiconductors?

Semiconductor materials are generally classified on the basis of the periodic table group that their constituent elements belong to. Thus, Group 12-16 semiconductors, formerly called II-VI semiconductors, are materials whose cations are from the Group 12 and anions are from Group 16 in the periodic table ([\[link\]](#)). Some examples of Group 12-16 semiconductor materials are cadmium selenide (CdSe), zinc sulfide (ZnS), cadmium telluride (CdTe), zinc oxide (ZnO), and mercuric selenide (HgSe) among others.

Note: The new IUPAC (International Union of Pure and Applied Chemistry) convention is being followed in this document, to avoid any confusion with regard to conventions used earlier. In the old IUPAC convention, Group 12 was known as Group IIB with the roman numeral 'II' referring to the number of electrons in the outer electronic shells and B referring to being on the right part of the table. However, in the CAS (Chemical Abstracts Service), the alphabet B refers to transition elements as compared to main group elements, though the roman numeral has the same meaning. Similarly, Group 16 was earlier known as Group VI because all the elements in this group have 6 valence shell electrons.

Group →	12	13	14	15	16
↓ Period					
2		5 B	6 C	7 N	8 O
3		13 Al	14 Si	15 P	16 S
4	30 Zn	31 Ga	32 Ge	33 As	34 Se
5	48 Cd	49 In	50 Sn	51 Sb	52 Te
6	80 Hg	81 Tl	82 Pb	83 Bi	84 Po
7	112 Cn	113 Uut	114 Uuq	115 Uup	116 Uuh

The red box
indicates the Group
12 and Group 16
elements in the
periodic table.

What are Group 12-16 (II-VI) semiconductor nanoparticles?

From the Greek word *nanos* - meaning "dwarf" this prefix is used in the metric system to mean 10^{-9} or one billionth ($1/1,000,000,000$). Thus a nanometer is 10^{-9} or one billionth of a meter, and a nanojoule is 10^{-9} or one billionth of a Joule, etc. A nanoparticle is ordinarily defined as any particle with at least one of its dimensions in the 1 - 100 nm range.

Nanoscale materials often show behavior which is intermediate between that of a bulk solid and that of an individual molecule or atom. An inorganic nanocrystal can be imagined to be comprised of a few atoms or molecules. It thus will behave differently from a single atom; however, it is still smaller than a macroscopic solid, and hence will show different properties. For example, if one would compare the chemical reactivity of a bulk solid and a nanoparticle, the latter would have a higher reactivity due to a significant fraction of the total number of atoms being on the surface of the particle. Properties such as boiling point, melting point, optical properties, chemical stability, electronic properties, etc. are all different in a nanoparticle as compared to its bulk counterpart. In the case of Group 12-16 semiconductors, this reduction in size from bulk to the nanoscale results in many size dependent properties such as varying band gap energy, optical and electronic properties.

Optical properties of semiconductor quantum nanoparticles

In the case of semiconductor nanocrystals, the effect of the size on the optical properties of the particles is very interesting. Consider a Group 12-16 semiconductor, cadmium selenide (CdSe). A 2 nm sized CdSe crystal

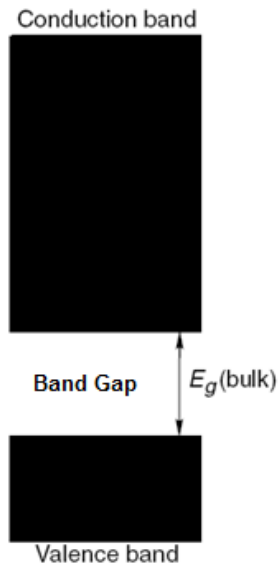
has a blue color fluorescence whereas a larger nanocrystal of CdSe of about 6 nm has a dark red fluorescence ([\[link\]](#)). In order to understand the size dependent optical properties of semiconductor nanoparticles, it is important to know the physics behind what is happening at the nano level.



Fluorescing CdSe quantum dots synthesized in a heat transfer liquid of different sizes (M. S. Wong, Rice University).

Energy levels in a semiconductor

The electronic structure of any material is given by a solution of Schrödinger equations with boundary conditions, depending on the physical situation. The electronic structure of a semiconductor ([\[link\]](#)) can be described by the following terms:



Simplified
representation
of the energy
levels in a
bulk
semiconductor

.

Energy level

By the solution of Schrödinger's equations, the electrons in a semiconductor can have only certain allowable energies, which are associated with energy levels. No electrons can exist in between these levels, or in other words can have energies in between the allowed energies. In addition, from Pauli's Exclusion Principle, only 2 electrons with opposite spin can exist at any one energy level. Thus, the electrons start filling from the lowest energy levels. Greater the number of atoms in a crystal, the difference in allowable energies become very small, thus the distance between energy levels decreases. However, this distance can never be zero. For a bulk semiconductor, due to the large number of atoms, the distance

between energy levels is very small and for all practical purpose the energy levels can be described as continuous ([\[link\]](#)).

Band gap

From the solution of Schrödinger's equations, there are a set of energies which is not allowable, and thus no energy levels can exist in this region. This region is called the band gap and is a quantum mechanical phenomenon ([\[link\]](#)). In a bulk semiconductor the bandgap is fixed; whereas in a quantum dot nanoparticle the bandgap varies with the size of the nanoparticle.

Valence band

In bulk semiconductors, since the energy levels can be considered as continuous, they are also termed as energy bands. Valence band contains electrons from the lowest energy level to the energy level at the lower edge of the bandgap ([\[link\]](#)). Since filling of energy is from the lowest energy level, this band is usually almost full.

Conduction band

The conduction band consists of energy levels from the upper edge of the bandgap and higher ([\[link\]](#)). To reach the conduction band, the electrons in the valence band should have enough energy to cross the band gap. Once the electrons are excited, they subsequently relax back to the valence band (either radiatively or non-radiatively) followed by a subsequent emission of radiation. This property is responsible for most of the applications of quantum dots.

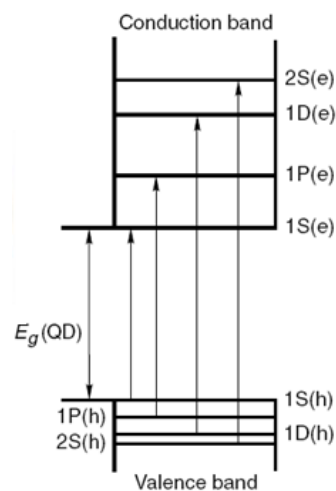
Exciton and exciton Bohr radius

When an electron is excited from the valence band to the conduction band, corresponding to the electron in the conduction band a hole (absence of electron) is formed in the valence band. This electron pair is called an exciton. Excitons have a natural separation distance between the electron and hole, which is characteristic of the material. This average distance is called exciton Bohr radius. In a bulk semiconductor, the size of the crystal is much larger than the exciton Bohr radius and hence the exciton is free to move throughout the crystal.

Energy levels in a quantum dot semiconductor

Before understanding the electronic structure of a quantum dot semiconductor, it is important to understand what a quantum dot nanoparticle is. We earlier studied that a nanoparticle is any particle with one of its dimensions in the 1 - 100 nm. A quantum dot is a nanoparticle with its diameter on the order of the materials exciton Bohr radius.

Quantum dots are typically 2 - 10 nm wide and approximately consist of 10 to 50 atoms. With this understanding of a quantum dot semiconductor, the electronic structure of a quantum dot semiconductor can be described by the following terms.



Energy levels in

quantum dot.
Allowed optical
transitions are
shown. Adapted
from T. Pradeep,
*Nano: The
Essentials.
Understanding
Nanoscience and
Nanotechnology*,
Tata McGraw-Hill,
New Delhi (2007).

Quantum confinement

When the size of the semiconductor crystal becomes comparable or smaller than the exciton Bohr radius, the quantum dots are in a state of quantum confinement. As a result of quantum confinement, the energy levels in a quantum dot are discrete ([\[link\]](#)) as opposed to being continuous in a bulk crystal ([\[link\]](#)).

Discrete energy levels

In materials that have small number of atoms and are considered as quantum confined, the energy levels are separated by an appreciable amount of energy such that they are not continuous, but are discrete (see [\[link\]](#)). The energy associated with an electron (equivalent to conduction band energy level) is given by $E_n = \frac{h^2 n^2}{8 m_e L^2}$, where h is the Planck's constant, m_e is the effective mass of electron and n is the quantum number for the conduction band states, and n can take the values 1, 2, 3 and so on. Similarly, the energy associated with the hole (equivalent to valence band energy level) is given by $E_{n'} = \frac{h^2 n'^2}{8 m_h L^2}$, where n' is the quantum number for the

valence states, and n' can take the values 1, 2, 3, and so on. The energy increases as one goes higher in the quantum number. Since the electron mass is much smaller than that of the hole, the electron levels are separated more widely than the hole levels.

Equation:

$$E^e = \frac{h^2 n^2}{8\pi^2 m_e d^2}$$

Equation:

$$E^h = \frac{h^2 n'^2}{8\pi^2 m_h d^2}$$

Tunable band gap

As seen from [\[link\]](#) and [\[link\]](#), the energy levels are affected by the diameter of the semiconductor particles. If the diameter is very small, since the energy is dependent on inverse of diameter squared, the energy levels of the upper edge of the band gap (lowest conduction band level) and lower edge of the band gap (highest valence band level) change significantly with the diameter of the particle and the effective mass of the electron and the hole, resulting in a size dependent tunable band gap. This also results in the discretization of the energy levels.

Qualitatively, this can be understood in the following way. In a bulk semiconductor, the addition or removal of an atom is insignificant compared to the size of the bulk semiconductor, which consists of a large number of atoms. The large size of bulk semiconductors makes the changes in band gap so negligible on the addition of an atom, that it is considered as a fixed band gap. In a quantum dot, addition of an atom does make a difference, resulting in the tunability of band gap.

UV-visible absorbance

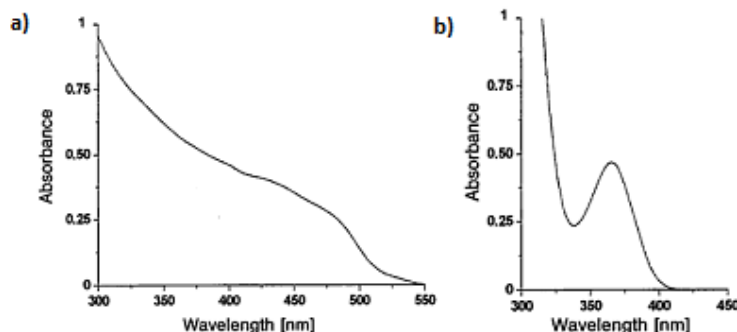
Due to the presence of discrete energy levels in a QD, there is a widening of the energy gap between the highest occupied electronic states and the lowest unoccupied states as compared to the bulk material. As a consequence, the optical properties of the semiconductor nanoparticles also become size dependent.

The minimum energy required to create an exciton is the defined by the band gap of the material, i.e., the energy required to excite an electron from the highest level of valence energy states to the lowest level of the conduction energy states. For a quantum dot, the bandgap varies with the size of the particle. From [\[link\]](#) and [\[link\]](#), it can be inferred that the band gap becomes higher as the particle becomes smaller. This means that for a smaller particle, the energy required for an electron to get excited is higher. The relation between energy and wavelength is given by [\[link\]](#), where h is the Planck's constant, c is the speed of light, λ is the wavelength of light. Therefore, from [\[link\]](#) to cross a bandgap of greater energy, shorter wavelengths of light are absorbed, i.e., a blue shift is seen.

Equation:

$$E = \frac{hc}{\lambda}$$

For Group 12-16 semiconductors, the bandgap energy falls in the UV-visible range. That is ultraviolet light or visible light can be used to excite an electron from the ground valence states to the excited conduction states. In a bulk semiconductor the band gap is fixed, and the energy states are continuous. This results in a rather uniform absorption spectrum ([\[link\]](#)a).



UV-vis spectra of (a) bulk CdS and (b) 4 nm CdS. Adapted from G. Kickelbick, *Hybrid Materials: Synthesis, Characterization and Applications*, Wiley-VCH, Weinheim (2007).

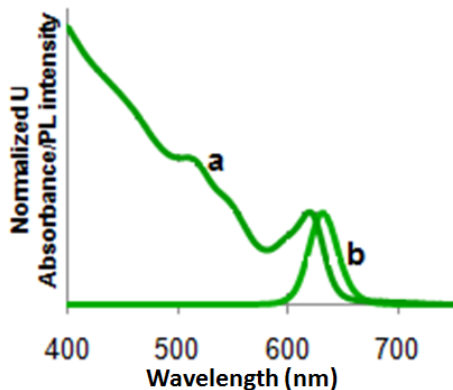
In the case of Group 12-16 quantum dots, since the bandgap can be changed with the size, these materials can absorb over a range of wavelengths. The peaks seen in the absorption spectrum ([link](#))b) correspond to the optical transitions between the electron and hole levels. The minimum energy and thus the maximum wavelength peak corresponds to the first exciton peak or the energy for an electron to get excited from the highest valence state to the lowest conduction state. The quantum dot will not absorb wavelengths of energy longer than this wavelength. This is known as the absorption onset.

Fluorescence

Fluorescence is the emission of electromagnetic radiation in the form of light by a material that has absorbed a photon. When a semiconductor quantum dot (QD) absorbs a photon/energy equal to or greater than its band gap, the electrons in the QD's get excited to the conduction state. This excited state is however not stable. The electron can relax back to its ground state by either emitting a photon or lose energy via heat losses. These processes can be divided into two categories – radiative decay and

non-radiative decay. Radiative decay is the loss of energy through the emission of a photon or radiation. Non-radiative decay involves the loss of heat through lattice vibrations and this usually occurs when the energy difference between the levels is small. Non-radiative decay occurs much faster than radiative decay.

Usually the electron relaxes to the ground state through a combination of both radiative and non-radiative decays. The electron moves quickly through the conduction energy levels through small non-radiative decays and the final transition across the band gap is via a radiative decay. Large nonradiative decays don't occur across the band gap because the crystal structure can't withstand large vibrations without breaking the bonds of the crystal. Since some of the energy is lost through the non-radiative decay, the energy of the emitted photon, through the radiative decay, is much lesser than the absorbed energy. As a result the wavelength of the emitted photon or fluorescence is longer than the wavelength of absorbed light. This energy difference is called the Stokes shift. Due this Stokes shift, the emission peak corresponding to the absorption band edge peak is shifted towards a higher wavelength (lower energy), i.e., [\[link\]](#).



Absorption spectra (a)
and emission spectra (b)
of CdSe tetrapod.

Intensity of emission versus wavelength is a bell-shaped Gaussian curve. As long as the excitation wavelength is shorter than the absorption onset, the maximum emission wavelength is independent of the excitation wavelength. [\[link\]](#) shows a combined absorption and emission spectrum for a typical CdSe tetrapod.

Factors affecting the optical properties of NPs

There are various factors that affect the absorption and emission spectra for Group 12-16 semiconductor quantum crystals. Fluorescence is much more sensitive to the background, environment, presence of traps and the surface of the QDs than UV-visible absorption. Some of the major factors influencing the optical properties of quantum nanoparticles include:

- **Surface defects, imperfection of lattice, surface charges** – The surface defects and imperfections in the lattice structure of semiconductor quantum dots occur in the form of unsatisfied valencies. Similar to surface charges, unsatisfied valencies provide a sink for the charge carriers, resulting in unwanted recombinations.
- **Surface ligands** – The presence of surface ligands is another factor that affects the optical properties. If the surface ligand coverage is a 100%, there is a smaller chance of surface recombinations to occur.
- **Solvent polarity** – The polarity of solvents is very important for the optical properties of the nanoparticles. If the quantum dots are prepared in organic solvent and have an organic surface ligand, the more non-polar the solvent, the particles are more dispersed. This reduces the loss of electrons through recombinations again, since when particles come in close proximity to each other, increases the non-radiative decay events.

Applications of the optical properties of Group 12-16 semiconductor NPs

The size dependent optical properties of NP's have many applications from biomedical applications to solar cell technology, from photocatalysis to chemical sensing. Most of these applications use the following unique properties.

For applications in the field of nanoelectronics, the sizes of the quantum dots can be tuned to be comparable to the scattering lengths, reducing the scattering rate and hence, the signal to noise ratio. For Group 12-16 QDs to be used in the field of solar cells, the bandgap of the particles can be tuned so as to form absorb energy over a large range of the solar spectrum, resulting in more number of excitons and hence more electricity. Since the nanoparticles are so small, most of the atoms are on the surface. Thus, the surface to volume ratio is very large for the quantum dots. In addition to a high surface to volume ratio, the Group 12-16 QDs respond to light energy. Thus quantum dots have very good photocatalytic properties. Quantum dots show fluorescence properties, and emit visible light when excited. This property can be used for applications as biomarkers. These quantum dots can be tagged to drugs to monitor the path of the drugs. Specially shaped Group 12-16 nanoparticles such as hollow shells can be used as drug delivery agents. Another use for the fluorescence properties of Group 12-16 semiconductor QDs is in color-changing paints, which can change colors according to the light source used.

Bibliography

- M. J. Schulz, V. N. Shanov, and Y. Yun, *Nanomedicine - Design of Particles, Sensors, Motors, Implants, Robots, and Devices*, Artech House, London (2009).
- S. V. Gapoenko, *Optical Properties of Semiconductor Nanocrystals*, Cambridge University Press, Cambridge (2003).
- T. Pradeep, *Nano: The Essentials. Understanding Nanoscience and Nanotechnology*, Tata McGraw-Hill, New Delhi (2007).
- G. Schmid, *Nanoparticles: From Theory to Application*, Wiley-VCH, Weinheim (2004).
- A. L. Rogach, *Semiconductor Nanocrystal Quantum Dots. Synthesis, Assembly, Spectroscopy and Applications*, Springer Wien, New York

(2008).

- G. Kickelbick, *Hybrid Materials: Synthesis, Characterization and Applications*, Wiley-VCH, Weinheim (2007).

Characterization of Group 12-16 (II-VI) Semiconductor Nanoparticles by UV-visible Spectroscopy

Quantum dots (QDs) as a general term refer to nanocrystals of semiconductor materials, in which the size of the particles are comparable to the natural characteristic separation of an electron-hole pair, otherwise known as the exciton Bohr radius of the material. When the size of the semiconductor nanocrystal becomes this small, the electronic structure of the crystal is governed by the laws of quantum physics. Very small Group 12-16 (II-VI) semiconductor nanoparticle quantum dots, in the order of 2 - 10 nm, exhibit significantly different optical and electronic properties from their bulk counterparts. The characterization of size dependent optical properties of Group 12-16 semiconductor particles provide a lot of qualitative and quantitative information about them – size, quantum yield, monodispersity, shape and presence of surface defects. A combination of information from both the UV-visible absorption and fluorescence, complete the analysis of the optical properties.

UV-visible absorbance spectroscopy

Absorption spectroscopy, in general, refers to characterization techniques that measure the absorption of radiation by a material, as a function of the wavelength. Depending on the source of light used, absorption spectroscopy can be broadly divided into infrared and UV-visible spectroscopy. The band gap of Group 12-16 semiconductors is in the UV-visible region. This means the minimum energy required to excite an electron from the valence states of the Group 12-16 semiconductor QDs to its conduction states, lies in the UV-visible region. This is also a reason why most of the Group 12-16 semiconductor quantum dot solutions are colored.

This technique is complementary to fluorescence spectroscopy, in that UV-visible spectroscopy measures electronic transitions from the ground state to the excited state, whereas fluorescence deals with the transitions from the excited state to the ground state. In order to characterize the optical properties of a quantum dot, it is important to characterize the sample with both these techniques

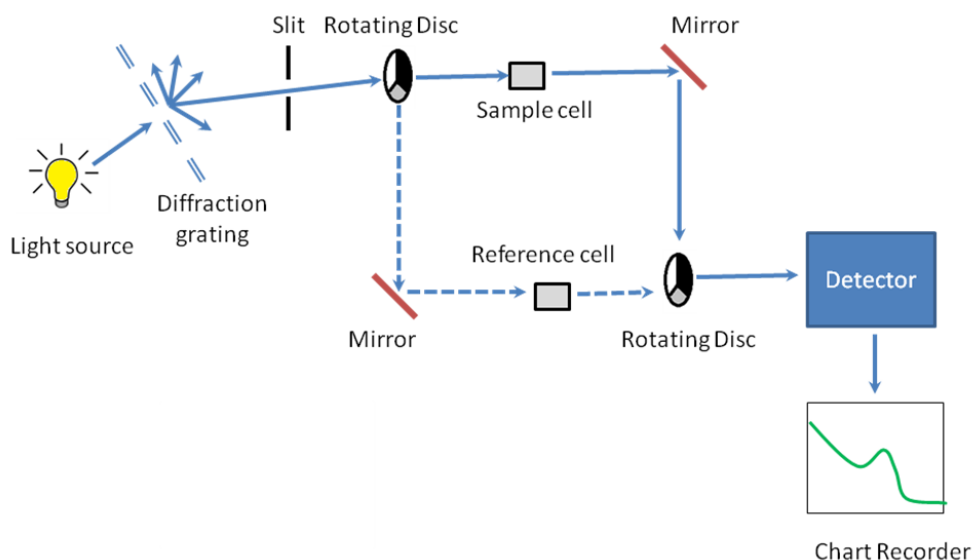
In quantum dots, due to the very small number of atoms, the addition or removal of one atom to the molecule changes the electronic structure of the quantum dot dramatically. Taking advantage of this property in Group 12-16 semiconductor quantum dots, it is possible to change the band gap of the material by just changing the size of the quantum dot. A quantum dot can absorb energy in the form of light over a range of wavelengths, to excite an electron from the ground state to its excited state. The minimum energy that is required to excite an electron, is dependent on the band gap of the quantum dot. Thus, by making accurate measurements of light absorption at different wavelengths in the ultraviolet and visible spectrum, a correlation can be made between the band gap and size of the quantum dot. Group 12-16 semiconductor quantum dots are of particular interest, since their band gap lies in the visible region of the solar spectrum.

The UV-visible absorbance spectroscopy is a characterization technique in which the absorbance of the material is studied as a function of wavelength. The visible region of the spectrum is in the wavelength range of 380 nm (violet) to 740 nm (red) and the near ultraviolet region extends to wavelengths of about 200 nm. The UV-visible spectrophotometer analyzes over the wavelength range 200 – 900 nm.

When the Group 12-16 semiconductor nanocrystals are exposed to light having an energy that matches a possible electronic transition as dictated by laws of quantum physics, the light is absorbed and an exciton pair is formed. The UV-visible spectrophotometer records the wavelength at which the absorption occurs along with the intensity of the absorption at each wavelength. This is recorded in a graph of absorbance of the nanocrystal versus wavelength.

Instrumentation

A working schematic of the UV-visible spectrophotometer is shown in [\[link\]](#).



Schematic of UV-visible spectrophotometer.

The light source

Since it is a UV-vis spectrophotometer, the light source ([\[link\]](#)) needs to cover the entire visible and the near ultra-violet region (200 - 900 nm). Since it is not possible to get this range of wavelengths from a single lamp, a combination of a deuterium lamp for the UV region of the spectrum and tungsten or halogen lamp for the visible region is used. This output is then sent through a diffraction grating as shown in the schematic.

The diffraction grating and the slit

The beam of light from the visible and/or UV light source is then separated into its component wavelengths (like a very efficient prism) by a diffraction grating ([\[link\]](#)). Following the slit is a slit that sends a monochromatic beam into the next section of the spectrophotometer.

Rotating discs

Light from the slit then falls onto a rotating disc ([\[link\]](#)). Each disc consists of different segments – an opaque black section, a transparent section and a mirrored section. If the light hits the transparent section, it will go straight through the sample cell, get reflected by a mirror, hits the mirrored section of a second rotating disc, and then collected by the detector. Else if the light hits the mirrored section, gets reflected by a mirror, passes through the reference cell, hits the transparent section of a second rotating disc and then collected by the detector. Finally if the light hits the black opaque section, it is blocked and no light passes through the instrument, thus enabling the system to make corrections for any current generated by the detector in the absence of light.

Sample cell, reference cell and sample preparation

For liquid samples, a square cross section tube sealed at one end is used. The choice of cuvette depends on the following factors:

- **Type of solvent** - For aqueous samples, specially designed rectangular quartz, glass or plastic cuvettes are used. For organic samples glass and quartz cuvettes are used.
- **Excitation wavelength** – Depending on the size and thus, bandgap of the 12-16 semiconductor nanoparticles, different excitation wavelengths of light are used. Depending on the excitation wavelength, different materials are used

Cuvette	Wavelength (nm)
Visible only glass	380 - 780

Visible only plastic	380 - 780
UV plastic	220 - 780
Quartz	200 - 900

Cuvette materials and their wavelengths.

- **Cost** – Plastic cuvettes are the least expensive and can be discarded after use. Though quartz cuvettes have the maximum utility, they are the most expensive, and need to be reused. Generally, disposable plastic cuvettes are used when speed is more important than high accuracy.

The best cuvettes need to be very clear and have no impurities that might affect the spectroscopic reading. Defects on the cuvette such as scratches, can scatter light and hence should be avoided. Some cuvettes are clear only on two sides, and can be used in the UV-Visible spectrophotometer, but cannot be used for fluorescence spectroscopy measurements. For Group 12-16 semiconductor nanoparticles prepared in organic solvents, the quartz cuvette is chosen.

In the sample cell the quantum dots are dispersed in a solvent, whereas in the reference cell the pure solvent is taken. It is important that the sample be very dilute (maximum first exciton absorbance should not exceed 1 au) and the solvent is not UV-visible active. For these measurements, it is required that the solvent does not have characteristic absorption or emission in the region of interest. Solution phase experiments are preferred, though it is possible to measure the spectra in the solid state also using thin films, powders, etc. The instrumentation for solid state UV-visible absorption spectroscopy is slightly different from the solution phase experiments and is beyond the scope of discussion.

Detector

Detector converts the light into a current signal that is read by a computer. Higher the current signal, greater is the intensity of the light. The computer

then calculates the absorbance using the in [\[link\]](#), where A denotes absorbance, I is sample cell intensity and I_0 is the reference cell intensity.

Equation:

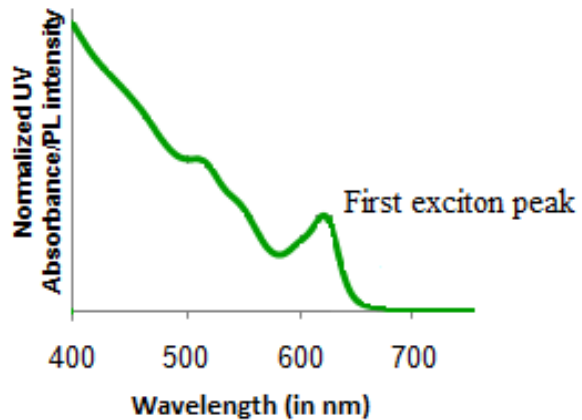
$$A = \log_{10}(I_0/I)$$

The following cases are possible:

- Where $I < I_0$ and $A < 0$. This usually occurs when the solvent absorbs in the wavelength range. Preferably the solvent should be changed, to get an accurate reading for actual reference cell intensity.
- Where $I = I_0$ and $A = 0$. This occurs when pure solvent is put in both reference and sample cells. This test should always be done before testing the sample, to check for the cleanliness of the cuvettes.
- When $A = 1$. This occurs when 90% of the light at a particular wavelength has been absorbed, which means that only 10% is seen at the detector. So I_0/I becomes $100/10 = 10$. \log_{10} of 10 is 1.
- When $A > 1$. This occurs in extreme case where more than 90% of the light is absorbed.

Output

The output is the form of a plot of absorbance against wavelength, e.g., [\[link\]](#).



Representative UV-visble absorption spectrum for CdSe tetrapods.

Beer-Lambert law

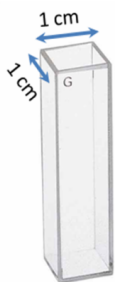
In order to make comparisons between different samples, it is important that all the factors affecting absorbance should be constant except the sample itself.

Effect of concentration on absorbance

The extent of absorption depends on the number of absorbing nanoparticles or in other words the concentration of the sample. If it is a reasonably concentrated solution, it will have a high absorbance since there are lots of nanoparticles to interact with the light. Similarly in an extremely dilute solution, the absorbance is very low. In order to compare two solutions, it is important that we should make some allowance for the concentration.

Effect of container shape

Even if we had the same concentration of solutions, if we compare two solutions – one in a rectangular shaped container (e.g., [link](#)) so that light travelled 1 cm through it and the other in which the light travelled 100 cm through it, the absorbance would be different. This is because if the length the light travelled is greater, it means that the light interacted with more number of nanocrystals, and thus has a higher absorbance. Again, in order to compare two solutions, it is important that we should make some allowance for the concentration.



A typical rectangular cuvette for UV-visible spectroscopy

The law

The Beer-Lambert law addresses the effect of concentration and container shape as shown in [link](#), [link](#) and [link](#), where A denotes absorbance; ϵ is the molar absorptivity or molar absorption coefficient; l is the path length of light (in cm); and c is the concentration of the solution (mol/dm^3).

Equation:

$$\log_{10}(I_0/I) = \epsilon lc$$

Equation:

$$A = \epsilon lc$$

Molar absorptivity

From the Beer-Lambert law, the molar absorptivity ' ϵ ' can be expressed as shown in [\[link\]](#).

Equation:

$$c = A/\epsilon l$$

Molar absorptivity corrects for the variation in concentration and length of the solution that the light passes through. It is the value of absorbance when light passes through 1 cm of a 1 mol/dm³ solution.

Limitations of Beer-Lambert law

The linearity of the Beer-Lambert law is limited by chemical and instrumental factors.

- At high concentrations (> 0.01 M), the relation between absorptivity coefficient and absorbance is no longer linear. This is due to the electrostatic interactions between the quantum dots in close proximity.
- If the concentration of the solution is high, another effect that is seen is the scattering of light from the large number of quantum dots.
- The spectrophotometer performs calculations assuming that the refractive index of the solvent does not change significantly with the presence of the quantum dots. This assumption only works at low concentrations of the analyte (quantum dots).

- Presence of stray light.

Analysis of data

The data obtained from the spectrophotometer is a plot of absorbance as a function of wavelength. Quantitative and qualitative data can be obtained by analysing this information

Quantitative Information

The band gap of the semiconductor quantum dots can be tuned with the size of the particles. The minimum energy for an electron to get excited from the ground state is the energy to cross the band gap. In an absorption spectra, this is given by the first exciton peak at the maximum wavelength (λ_{max}).

Size of the quantum dots

The size of quantum dots can be approximated corresponding to the first exciton peak wavelength. Empirical relationships have been determined relating the diameter of the quantum dot to the wavelength of the first exciton peak. The Group 12-16 semiconductor quantum dots that they studied were cadmium selenide (CdSe), cadmium telluride (CdTe) and cadmium sulfide (CdS). The empirical relationships are determined by fitting experimental data of absorbance versus wavelength of known sizes of particles. The empirical equations determined are given for CdTe, CdSe, and CdS in [\[link\]](#), [\[link\]](#) and [\[link\]](#) respectively, where D is the diameter and λ is the wavelength corresponding to the first exciton peak. For example, if the first exciton peak of a CdSe quantum dot is 500 nm, the corresponding diameter of the quantum dot is 2.345 nm and for a wavelength of 609 nm, the corresponding diameter is 5.008 nm.

Equation:

$$D = (9.8127 \times 10^{-7})\lambda^3 - (1.7147 \times 10^{-3})\lambda^2 + (1.0064)\lambda - 194.84$$

Equation:

$$D = (1.6122 \times 10^{-7})\lambda^3 - (2.6575 \times 10^{-6})\lambda^2 + (1.6242 \times 10^{-3})\lambda + 41.57$$

Equation:

$$D = (-6.6521 \times 10^{-8})\lambda^3 + (1.9577 \times 10^{-4})\lambda^2 - (9.2352 \times 10^{-2})\lambda + 13.29$$

Concentration of sample

Using the Beer-Lambert law, it is possible to calculate the concentration of the sample if the molar absorptivity for the sample is known. The molar absorptivity can be calculated by recording the absorbance of a standard solution of 1 mol/dm³ concentration in a standard cuvette where the light travels a constant distance of 1 cm. Once the molar absorptivity and the absorbance of the sample are known, with the length the light travels being fixed, it is possible to determine the concentration of the sample solution.

Empirical equations can be determined by fitting experimental data of extinction coefficient per mole of Group 12-16 semiconductor quantum dots, at 250 °C, to the diameter of the quantum dot, [\[link\]](#), [\[link\]](#), and [\[link\]](#).

Equation:

$$\varepsilon = 10043 \times D^{2.12}$$

Equation:

$$\varepsilon = 5857 \times D^{2.65}$$

Equation:

$$\varepsilon = 21536 \times D^{2.3}$$

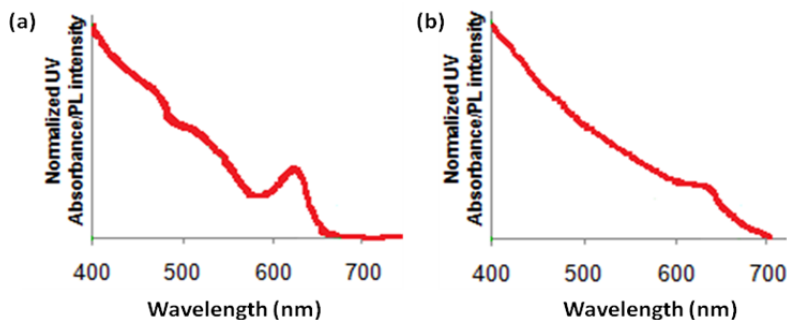
The concentration of the quantum dots can then be then be determined by using the Beer Lambert law as given by [\[link\]](#).

Qualitative Information

Apart from quantitative data such as the size of the quantum dots and concentration of the quantum dots, a lot of qualitative information can be derived from the absorption spectra.

Size distribution

If there is a very narrow size distribution, the first exciton peak will be very sharp ([\[link\]](#)). This is because due to the narrow size distribution, the differences in band gap between different sized particles will be very small and hence most of the electrons will get excited over a smaller range of wavelengths. In addition, if there is a narrow size distribution, the higher exciton peaks are also seen clearly.

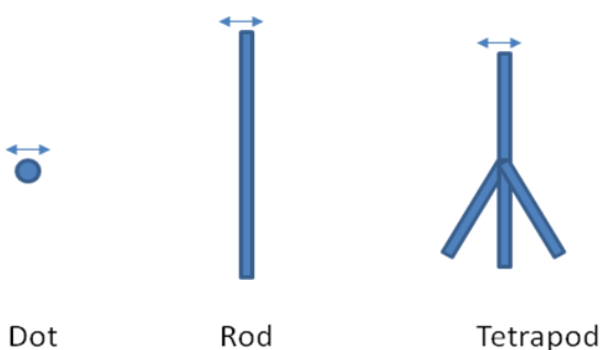


Narrow emission spectra (a) and broad emission spectra (b) of CdSe QDs.

Shaped particles

In the case of a spherical quantum dot, in all dimensions, the particle is quantum confined ([\[link\]](#)). In the case of a nanorod, whose length is not in the quantum regime, the quantum effects are determined by the width of the

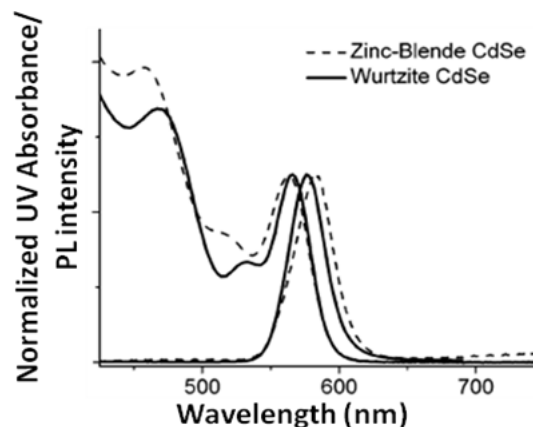
nanorod. Similar is the case in tetrapods or four legged structures. The quantum effects are determined by the thickness of the arms. During the synthesis of the shaped particles, the thickness of the rod or the arm of the tetrapod does not vary among the different particles, as much as the length of the rods or arms changes. Since the thickness of the rod or tetrapod is responsible for the quantum effects, the absorption spectrum of rods and tetrapods has sharper features as compared to a quantum dot. Hence, qualitatively it is possible to differentiate between quantum dots and other shaped particles.



Different shaped nanoparticles with the arrows indicating the dimension where quantum confinement effects are observed.

Crystal lattice information

In the case of CdSe semiconductor quantum dots it has been shown that it is possible to estimate the crystal lattice of the quantum dot from the adsorption spectrum ([\[link\]](#)), and hence determine if the structure is zinc blend or wurtzite.



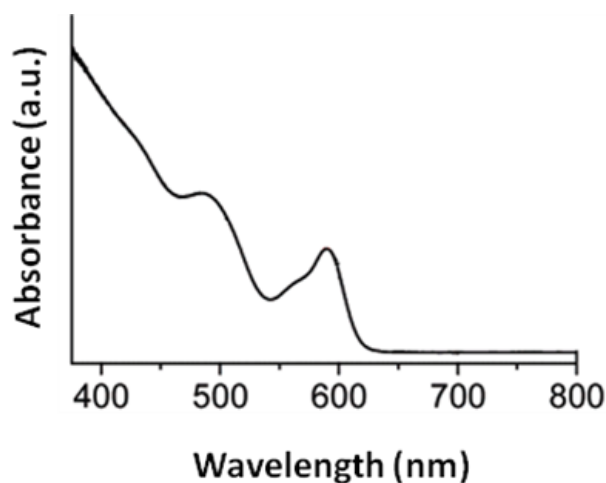
Zinc blende and wurtzite CdSe absorption spectra. Adapted from J. Jasieniak, C. Bullen, J. van Embden, and P. Mulvaney, *J. Phys. Chem. B*, 2005, **109**, 20665.

UV-vis absorption spectra of Group 12-16 semiconductor nanoparticles

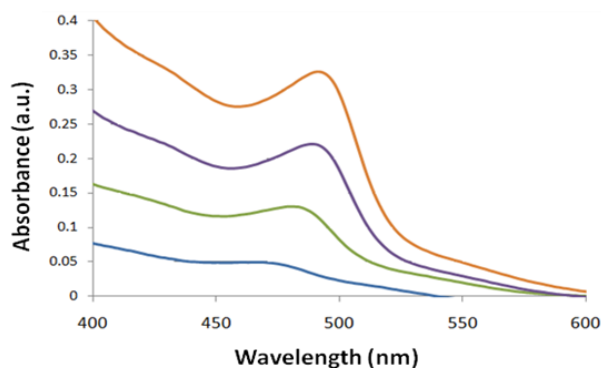
Cadmium selenide

Cadmium selenide (CdSe) is one of the most popular Group 12-16 semiconductors. This is mainly because the band gap (712 nm or 1.74 eV) energy of CdSe. Thus, the nanoparticles of CdSe can be engineered to have a range of band gaps throughout the visible range, corresponding to the major part of the energy that comes from the solar spectrum. This property of CdSe along with its fluorescing properties is used in a variety of applications such as solar cells and light emitting diodes. Though cadmium and selenium are known carcinogens, the harmful biological effects of CdSe can be overcome by coating the CdSe with a layer of zinc sulfide. Thus CdSe, can also be used as bio-markers, drug-delivery agents, paints and other applications.

A typical absorption spectrum of narrow size distribution wurtzite CdSe quantum dot is shown in [\[link\]](#). A size evolving absorption spectra is shown in [\[link\]](#). However, a complete analysis of the sample is possible only by also studying the fluorescence properties of CdSe.



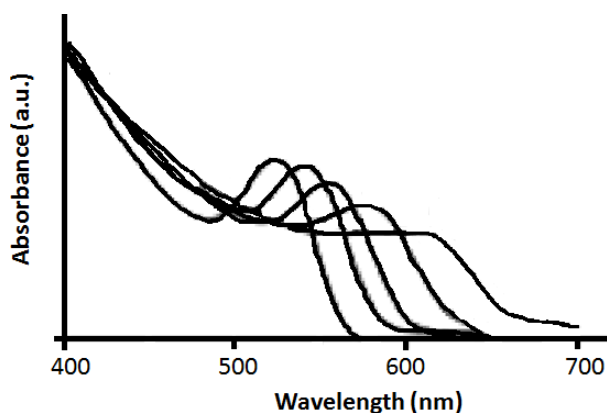
Wurtzite CdSe quantum dot.
Adapted from X. Zhong, Y.
Feng, and Y. Zhang, *J. Phys.*
Chem. C, 2007, **111**, 526.



Size evolving absorption
spectra of CdSe quantum dots.

Cadmium telluride (CdTe)

Cadmium telluride has a band gap of 1.44 eV (860 nm) and as such it absorbs in the infrared region. Like CdSe, the sizes of CdTe can be engineered to have different band edges and thus, different absorption spectra as a function of wavelength. A typical CdTe spectra is shown in [\[link\]](#). Due to the small bandgap energy of CdTe, it can be used in tandem with CdSe to absorb in a greater part of the solar spectrum.



Size evolving absorption spectra of CdTe quantum dots from 3 nm to 7 nm. Adapted from C. Qi-Fan, W. Wen-Xing, G. Ying-Xin, L. Meng-Ying, X. Shu-Kun and Z. Xiu-Juan, *Chin. J. Anal. Chem.*, 2007, **35**, 135.

Other Group 12-16 semiconductor systems

[\[link\]](#) shows the bulk band gap of other Group 12-16 semiconductor systems. The band gap of ZnS falls in the UV region, while those of ZnSe, CdS, and ZnTe fall in the visible region.

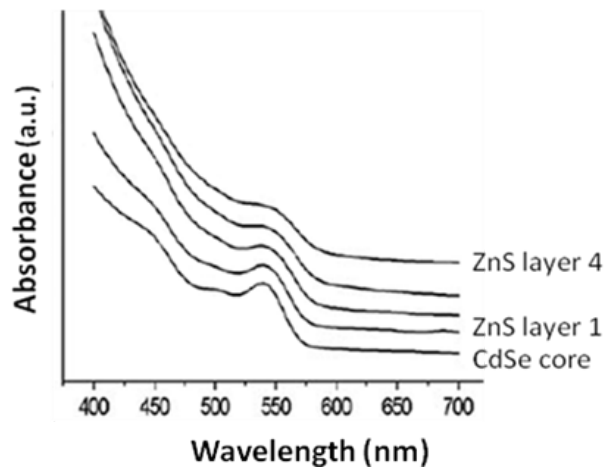
Material	Band gap (eV)	Wavelength (nm)
ZnS	3.61	343.2
ZnSe	2.69	460.5
ZnTe	2.39	518.4
CdS	2.49	497.5
CdSe	1.74	712.1
CdTe	1.44	860.3

Bulk band gaps of different Group 12-16 semiconductors.

Heterostructures of Group 12-16 semiconductor systems

It is often desirable to have a combination of two Group 12-16 semiconductor system quantum heterostructures of different shapes like dots and tetrapods, for applications in solar cells, bio-markers, etc. Some of the most interesting systems are ZnS shell-CdSe core systems, such as the CdSe/CdS rods and tetrapods.

[\[link\]](#) shows a typical absorption spectra of CdSe-ZnS core-shell system. This system is important because of the drastically improved fluorescence properties because of the addition of a wide band gap ZnS shell than the core CdSe. In addition with a ZnS shell, CdSe becomes bio-compatible.



Absorption spectra of CdSe core, ZnS shell. Adapted from C. Qing-Zhu, P. Wang, X. Wang and Y. Li, *Nanoscale Res. Lett.*, 2008, 3, 213.

A CdSe seed, CdS arm nanorods system is also interesting. Combining CdSe and CdS in a single nanostructure creates a material with a mixed dimensionality where holes are confined to CdSe while electrons can move freely between CdSe and CdS phases.

Bibliography

- S. V. Gapoenko, *Optical Properties of Semiconductor Nanocrystals*, Cambridge University Press, Cambridge (2003).
- W. W. Yu, L. Qu, W. Guo, and X. Peng, *Chem. Mater.*, 2003, 15, 2854.

- J. Jasieniak, C. Bullen, J. van Embden, and P. Mulvaney, *J. Phys. Chem. B*, 2005, **109**, 20665.
- X. Zhong, Y. Feng, and Y. Zhang, *J. Phys. Chem. C*, 2007, **111**, 526.
- D. V. Talapin, J. H. Nelson, E. V. Shevchenko, S. Aloni, B. Sadtler, and A. P. Alivisatos, *Nano Lett.*, 2007, **7**, 2951.
- C. Qing-Zhu, P. Wang, X. Wang, and Y. Li, *Nanoscale Res. Lett.*, 2008, **3**, 213.
- C. Qi-Fan, W. Wen-Xing, G. Ying-Xin, L. Meng-Ying, X. Shu-Kun, and Z. Xiu-Juan, *Chin. J. Anal. Chem.*, 2007, **35**, 135.

Optical Characterization of Group 12-16 (II-VI) Semiconductor Nanoparticles by Fluorescence Spectroscopy

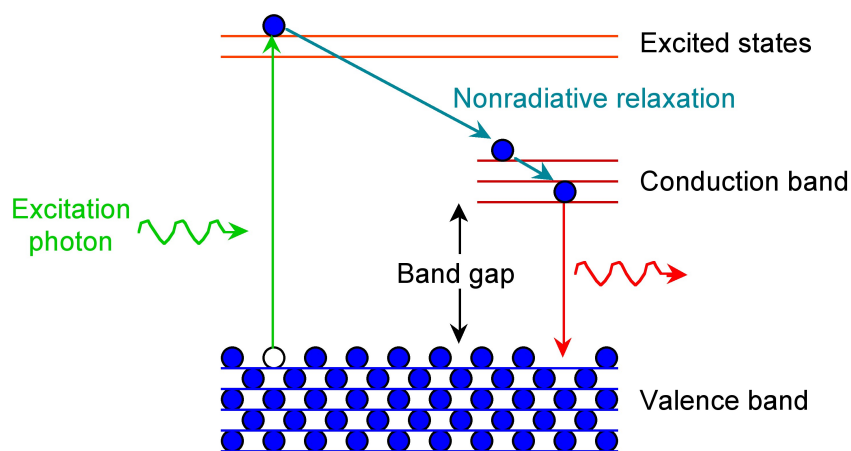
Group 12-16 semiconductor nanocrystals when exposed to light of a particular energy absorb light to excite electrons from the ground state to the excited state, resulting in the formation of an electron-hole pair (also known as excitons). The excited electrons relax back to the ground state, mainly through radiative emission of energy in the form of photons.

Quantum dots (QD) refer to nanocrystals of semiconductor materials where the size of the particles is comparable to the natural characteristic separation of an electron-hole pair, otherwise known as the exciton Bohr radius of the material. In quantum dots, the phenomenon of emission of photons associated with the transition of electrons from the excited state to the ground state is called fluorescence.

Fluorescence spectroscopy

Emission spectroscopy, in general, refers to a characterization technique that measures the emission of radiation by a material that has been excited. Fluorescence spectroscopy is one type of emission spectroscopy which records the intensity of light radiated from the material as a function of wavelength. It is a nondestructive characterization technique.

After an electron is excited from the ground state, it needs to relax back to the ground state. This relaxation or loss of energy to return to the ground state, can be achieved by a combination of non-radiative decay (loss of energy through heat) and radiative decay (loss of energy through light). Non-radiative decay by vibrational modes typically occurs between energy levels that are close to each other. Radiative decay by the emission of light occurs when the energy levels are far apart like in the case of the band gap. This is because loss of energy through vibrational modes across the band gap can result in breaking the bonds of the crystal. This phenomenon is shown in [\[link\]](#).



Emission of luminescence photon for Group 12-16 semiconductor quantum dot.

The band gap of Group 12-16 semiconductors is in the UV-visible region. Thus, the wavelength of the emitted light as a result of radiative decay is also in the visible region, resulting in fascinating fluorescence properties.

A fluorimeter is a device that records the fluorescence intensity as a function of wavelength. The fluorescence quantum yield can then be calculated by the ratio of photons absorbed to photons emitted by the system. The quantum yield gives the probability of the excited state getting relaxed via fluorescence rather than by any other non-radiative decay.

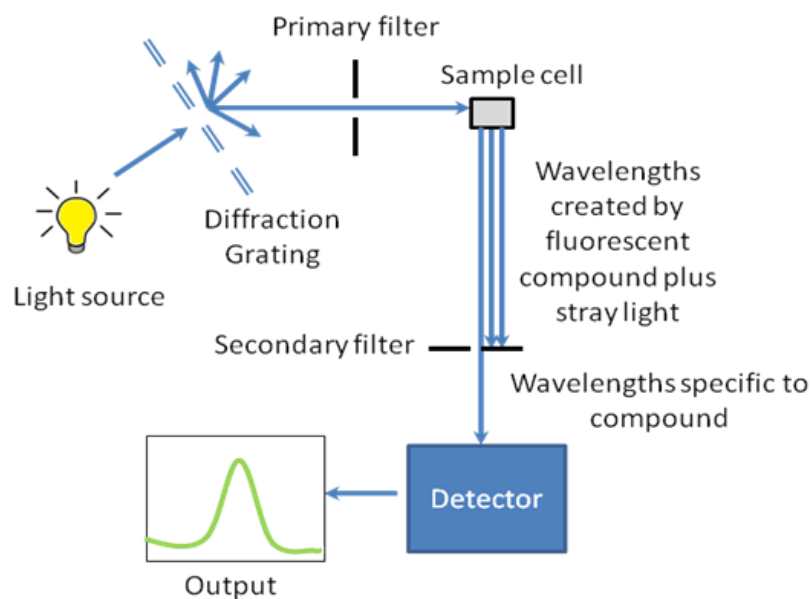
Difference between fluorescence and phosphorescence

Photoluminescence is the emission of light from any material due to the loss of energy from excited state to ground state. There are two main types of luminescence – fluorescence and phosphorescence. Fluorescence is a fast decay process, where the emission rate is around 10^8 s^{-1} and the lifetime is around $10^{-9} - 10^{-7} \text{ s}$. Fluorescence occurs when the excited state electron has an opposite spin compared to the ground state electrons. From the laws of quantum mechanics, this is an allowed transition, and occurs rapidly by emission of a photon. Fluorescence disappears as soon as the exciting light source is removed.

Phosphorescence is the emission of light, in which the excited state electron has the same spin orientation as the ground state electron. This transition is a forbidden one and hence the emission rates are slow ($10^3 - 10^0 \text{ s}^{-1}$). So the phosphorescence lifetimes are longer, typically seconds to several minutes, while the excited phosphors slowly returned to the ground state. Phosphorescence is still seen, even after the exciting light source is removed. Group 12-16 semiconductor quantum dots exhibit fluorescence properties when excited with ultraviolet light.

Instrumentation

The working schematic for the fluorometer is shown in [\[link\]](#).



Schematic of fluorometer.

The light source

The excitation energy is provided by a light source that can emit wavelengths of light over the ultraviolet and the visible range. Different light sources can be used as excitation sources such as lasers, xenon arcs and mercury-vapor lamps. The choice of the light source depends on the sample. A laser source emits light of a high irradiance at a very narrow wavelength interval. This makes the need for the filter unnecessary, but the wavelength of the laser cannot be altered significantly. The mercury vapor lamp is a discrete line source. The xenon arc has a continuous emission spectrum between the ranges of 300 - 800 nm.

The diffraction grating and primary filter

The diffraction grating splits the incoming light source into its component wavelengths ([\[link\]](#)). The monochromator can then be adjusted to choose with wavelengths to pass through. Following the primary filter, specific wavelengths of light are irradiated onto the sample

Sample cell and sample preparation

A proportion of the light from the primary filter is absorbed by the sample. After the sample gets excited, the fluorescent substance returns to the ground state, by emitting a longer wavelength of light in all directions ([\[link\]](#)). Some of this light passes through a secondary filter. For liquid samples, a square cross section tube sealed at one end and all four sides clear, is used as a sample cell. The choice of cuvette depends on three factors:

1. **Type of solvent** - For aqueous samples, specially designed rectangular quartz, glass or plastic cuvettes are used. For organic samples glass and quartz cuvettes are used.
2. **Excitation wavelength** – Depending on the size and thus, bandgap of the Group 12-16 semiconductor nanoparticles, different excitation wavelengths of light are used. Depending on the excitation wavelength, different materials are used ([\[link\]](#)).

Cuvette	Wavelength (nm)
Visible only glass	380 - 780
Visible only plastic	380 - 780
UV plastic	220 - 780
Quartz	200 - 900

Cuvette materials and their wavelengths.

3. **Cost** – Plastic cuvettes are the least expensive and can be discarded after use. Though quartz cuvettes have the maximum utility, they are the most expensive, and need to be reused. Generally, disposable plastic cuvettes are used when speed is more important than high accuracy.



A typical
cuvette for
fluorescence
spectroscopy

The cuvettes have a 1 cm path length for the light ([\[link\]](#)). The best cuvettes need to be very clear and have no impurities that might affect the

spectroscopic reading. Defects on the cuvette, such as scratches, can scatter light and hence should be avoided. Since the specifications of a cuvette are the same for both, the UV-visible spectrophotometer and fluorimeter, the same cuvette that is used to measure absorbance can be used to measure the fluorescence. For Group 12-16 semiconductor nanoparticles prepared in organic solvents, the clear four sided quartz cuvette is used. The sample solution should be dilute (absorbance <1 au), to avoid very high signal from the sample to burn out the detector. The solvent used to disperse the nanoparticles should not absorb at the excitation wavelength.

Secondary filter

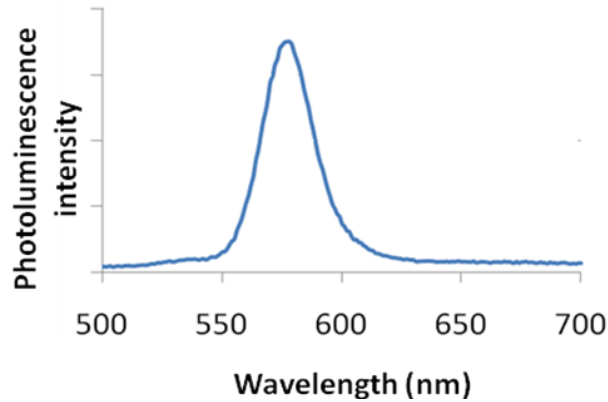
The secondary filter is placed at a 90° angle ([\[link\]](#)) to the original light path to minimize the risk of transmitted or reflected incident light reaching the detector. Also this minimizes the amount of stray light, and results in a better signal-to-noise ratio. From the secondary filter, wavelengths specific to the sample are passed onto the detector.

Detector

The detector can either be single-channeled or multichanneled ([\[link\]](#)). The single-channeled detector can only detect the intensity of one wavelength at a time, while the multichanneled detects the intensity at all wavelengths simultaneously, making the emission monochromator or filter unnecessary. The different types of detectors have both advantages and disadvantages.

Output

The output is the form of a plot of intensity of emitted light as a function of wavelength as shown in [\[link\]](#).



Emission spectra of CdSe quantum dot.

Analysis of data

The data obtained from fluorimeter is a plot of fluorescence intensity as a function of wavelength. Quantitative and qualitative data can be obtained by analysing this information.

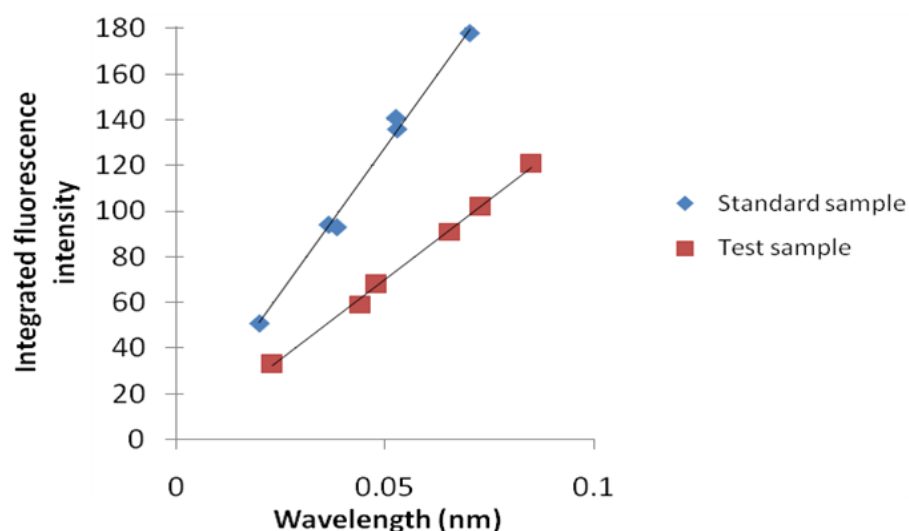
Quantitative information

From the fluorescence intensity versus wavelength data, the quantum yield (Φ_F) of the sample can be determined. Quantum yield is a measure of the ratio of the photons absorbed with respect to the photons emitted. It is important for the application of Group 12-16 semiconductor quantum dots using their fluorescence properties, for e.g., bio-markers.

The most well-known method for recording quantum yield is the comparative method which involves the use of well characterized standard solutions. If a test sample and a standard sample have similar absorbance values at the same excitation wavelength, it can be assumed that the number of photons being absorbed by both the samples is the same. This means that a ratio of the integrated fluorescence intensities of the test and standard

sample measured at the same excitation wavelength will give a ratio of quantum yields. Since the quantum yield of the standard solution is known, the quantum yield for the unknown sample can be calculated.

A plot of integrated fluorescence intensity versus absorbance at the excitation wavelength is shown in [\[link\]](#). The slope of the graphs shown in [\[link\]](#) are proportional to the quantum yield of the different samples. Quantum yield is then calculated using [\[link\]](#), where subscripts ST denotes standard sample and X denotes the test sample; QY is the quantum yield; RI is the refractive index of the solvent.



Integrated fluoresncene intensity as a function of absorbance.

Equation:

$$\frac{QY_X}{QY_{ST}} = \frac{\text{slope}_X (RI_X)^2}{\text{slope}_{ST} (RI_{ST})^2}$$

Take the example of [\[link\]](#). If the same solvent is used in both the sample and the standard solution, the ratio of quantum yields of the sample to the

standard is given by [\[link\]](#). If the quantum yield of the standard is known to 0.95, then the quantum yield of the test sample is 0.523 or 52.3%.

Equation:

$$\frac{QY_X}{QY_{ST}} = \frac{1.41}{2.56}$$

The assumption used in the comparative method is valid only in the Beer-Lambert law linear regime. Beer-Lambert law states that absorbance is directly proportional to the path length of light travelled within the sample, and concentration of the sample. The factors that affect the quantum yield measurements are the following:

- **Concentration** – Low concentrations should be used (absorbance < 0.2 a.u.) to avoid effects such as self quenching.
- **Solvent** – It is important to take into account the solvents used for the test and standard solutions. If the solvents used for both are the same then the comparison is trivial. However, if the solvents in the test and standard solutions are different, this difference needs to be accounted for. This is done by incorporating the solvent refractive indices in the ratio calculation.
- **Standard samples** – The standard samples should be characterized thoroughly. In addition, the standard sample used should absorb at the excitation wavelength of the test sample.
- **Sample preparation** – It is important that the cuvettes used are clean, scratch free and clear on all four sides. The solvents used must be of spectroscopic grade and should not absorb in the wavelength range.
- **Slit width** – The slit widths for all measurements must be kept constant.

The quantum yield of the Group 12-16 semiconductor nanoparticles are affected by many factors such as the following.

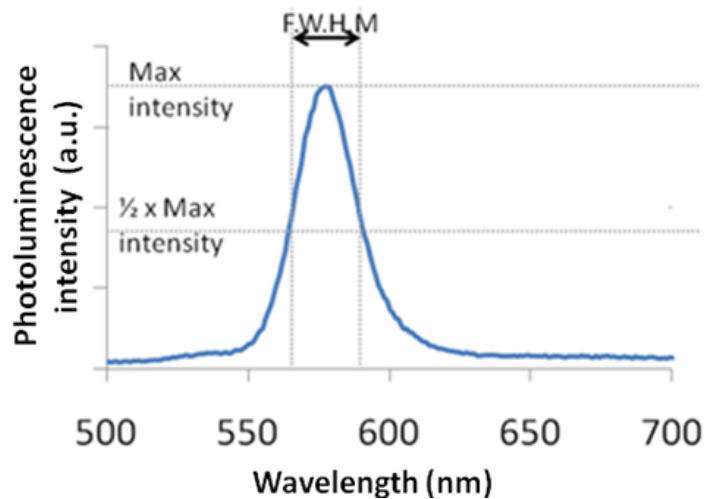
- **Surface defects** – The surface defects of semiconductor quantum dots occur in the form of unsatisfied valencies. Thus resulting in unwanted recombinations. These unwanted recombinations reduce the loss of energy through radiative decay, and thus reducing the fluorescence.

- **Surface ligands** – If the surface ligand coverage is a 100%, there is a smaller chance of surface recombinations to occur.
- **Solvent polarity** – If the solvent and the ligand have similar solvent polarities, the nanoparticles are more dispersed, reducing the loss of electrons through recombinations.

Qualitative Information

Apart from quantum yield information, the relationship between intensity of fluorescence emission and wavelength, other useful qualitative information such as size distribution, shape of the particle and presence of surface defects can be obtained.

As shown in [\[link\]](#), the shape of the plot of intensity versus wavelength is a Gaussian distribution. In [\[link\]](#), the full width at half maximum (FWHM) is given by the difference between the two extreme values of the wavelength at which the photoluminescence intensity is equal to half its maximum value. From the full width half max (FWHM) of the fluorescence intensity Gaussian distribution, it is possible to determine qualitatively the size distribution of the sample. For a Group 12-16 quantum dot sample if the FWHM is greater than 30, the system is very polydisperse and has a large size distribution. It is desirable for all practical applications for the FWHM to be lesser than 30.



Emission spectra of CdSe QDs showing the full width half maximum (FWHM).

From the FWHM of the emission spectra, it is also possible to qualitatively get an idea if the particles are spherical or shaped. During the synthesis of the shaped particles, the thickness of the rod or the arm of the tetrapod does not vary among the different particles, as much as the length of the rods or arms changes. The thickness of the arm or rod is responsible for the quantum effects in shaped particles. In the case of quantum dots, the particle is quantum confined in all dimensions. Thus, any size distribution during the synthesis of quantum dots greatly affects the emission spectra. As a result the FWHM of rods and tetrapods is much smaller as compared to a quantum dot. Hence, qualitatively it is possible to differentiate between quantum dots and other shaped particles.

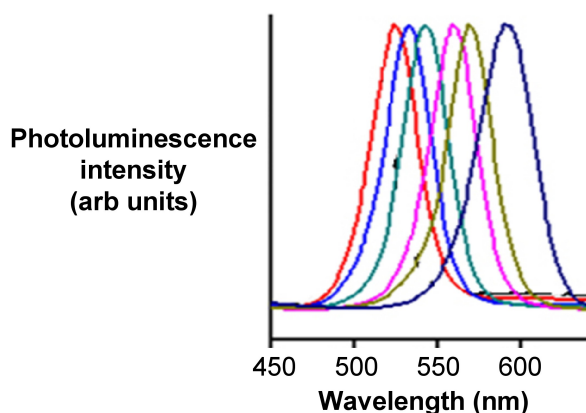
Another indication of branched structures is the decrease in the intensity of fluorescence peaks. Quantum dots have very high fluorescence values as compared to branched particles, since they are quantum confined in all dimensions as compared to just 1 or 2 dimensions in the case of branched particles.

Fluorescence spectra of different Group 12-16 semiconductor nanoparticles

The emission spectra of all Group 12-16 semiconductor nanoparticles are Gaussian curves as shown in [\[link\]](#) and [\[link\]](#). The only difference between them is the band gap energy, and hence each of the Group 12-16 semiconductor nanoparticles fluoresce over different ranges of wavelengths

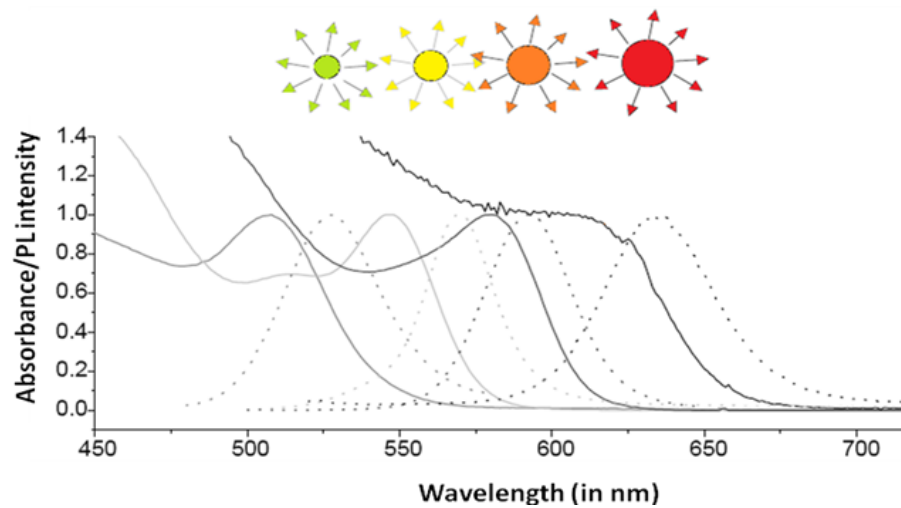
Cadmium selenide

Since its bulk band gap (1.74 eV, 712 nm) falls in the visible region cadmium Selenide (CdSe) is used in various applications such as solar cells, light emitting diodes, etc. Size evolving emission spectra of cadmium selenide is shown in [\[link\]](#). Different sized CdSe particles have different colored fluorescence spectra. Since cadmium and selenide are known carcinogens and being nanoparticles are easily absorbed into the human body, there is some concern regarding these particles. However, CdSe coated with ZnS can overcome all the harmful biological effects, making cadmium selenide nanoparticles one of the most popular 12-16 semiconductor nanoparticle.



Size evolving CdSe emission spectra. Adapted from <http://www.physics.mq.edu.au>.

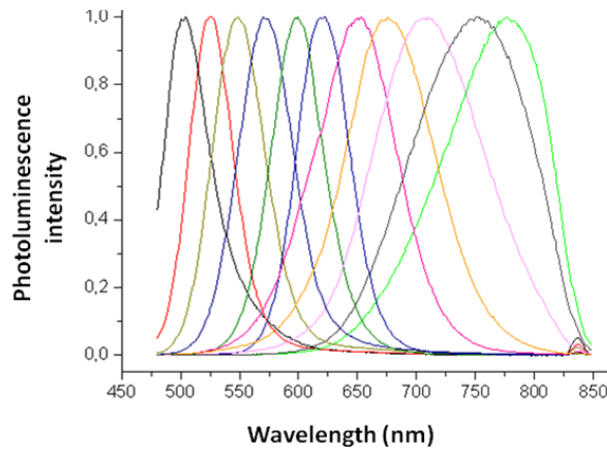
A combination of the absorbance and emission spectra is shown in [\[link\]](#) for four different sized particles emitting green, yellow, orange, and red fluorescence.



Absorption and emission spectra of CdSe quantum dots. Adapted from G. Schmid, *Nanoparticles: From Theory to Application*, Wiley-VCH, Weinham (2004).

Cadmium telluride

Cadmium Telluride (CdTe) has a band gap of 1.44 eV and thus absorbs in the infra red region. The size evolving CdTe emission spectra is shown in [\[link\]](#).

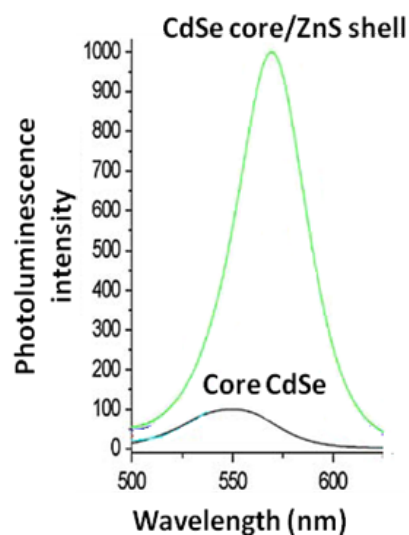


Size evolution spectra of CdTe quantum dots.

Adding shells to QDs

Capping a core quantum dot with a semiconductor material with a wider bandgap than the core, reduces the nonradiative recombination and results in brighter fluorescence emission. Quantum yields are affected by the presences of free surface charges, surface defects and crystal defects, which results in unwanted recombinations. The addition of a shell reduces the nonradiative transitions and majority of the electrons relax radiatively to the valence band. In addition, the shell also overcomes some of the surface defects.

For the CdSe-core/ZnS-shell systems exhibit much higher quantum yield as compared to core CdSe quantum dots as seen in [\[link\]](#).



Emission spectra of core CdSe only and CdSe-core/ZnS-shell.

Bibliography

- A. T. R. Williams, S. A. Winfield, and J. N. Miller, *Analyst*, 1983, **108**, 1067.
- G. Schmid, *Nanoparticles: From Theory to Application*, Wiley-VCH, Weinham, (2004).
- J. Y. Hariba, *A Guide to Recording Fluorescence Quantum Yield*, Jobin Yvon Hariba Limited, Stanmore (2003).
- C. Qing Zhu, P. Wang, X. Wang, and Y. Li, *Nanoscale Res. Lett.*, 2008, **3**, 213.

Carbon Nanomaterials

Introduction

Although nanomaterials had been known for many years prior to the report of C_{60} the field of nanoscale science was undoubtedly founded upon this seminal discovery. Part of the reason for this explosion in nanochemistry is that while carbon materials range from well-defined nano sized molecules (i.e., C_{60}) to tubes with lengths of hundreds of microns, they do not exhibit the instabilities of other nanomaterials as a result of the very high activation barriers to their structural rearrangement. As a consequence they are highly stable even in their unfunctionalized forms. Despite this range of carbon nanomaterials possible they exhibit common reaction chemistry: that of organic chemistry.

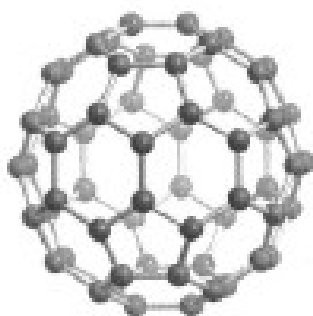
The previously unknown allotrope of carbon: C_{60} , was discovered in 1985, and in 1996, Curl, Kroto, and Smalley were awarded the Nobel Prize in Chemistry for the discovery. The other allotropes of carbon are graphite (sp^2) and diamond (sp^3). C_{60} , commonly known as the “buckyball” or “Buckminsterfullerene”, has a spherical shape comprising of highly pyramidalized sp^2 carbon atoms. The C_{60} variant is often compared to the typical soccer football, hence buckyball. However, confusingly, this term is commonly used for higher derivatives. Fullerenes are similar in sheet structure to graphite but they contain pentagonal (or sometimes heptagonal) rings that prevent the sheet from being planar. The unusual structure of C_{60} led to the introduction of a new class of molecules known as fullerenes, which now constitute the third allotrope of carbon. Fullerenes are commonly defined as “any of a class of closed hollow aromatic carbon compounds that are made up of twelve pentagonal and differing numbers of hexagonal faces.”

The number of carbon atoms in a fullerene range from C_{60} to C_{70} , C_{76} , and higher. Higher order fullerenes include carbon nanotubes that can be described as fullerenes that have been stretched along a rotational axis to form a tube. As a consequence of differences in the chemistry of fullerenes such as C_{60} and C_{70} as compared to nanotubes, these will be dealt with separately herein. In addition there have also been reports of nanohorns and

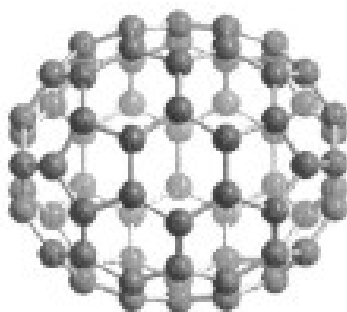
nanofibers, however, these may be considered as variations on the general theme. It should be noted that fullerenes and nanotubes have been shown to be in flames produced by hydrocarbon combustion. Unfortunately, these naturally occurring varieties can be highly irregular in size and quality, as well as being formed in mixtures, making them unsuitable for both research and industrial applications.

Fullerenes

Carbon-60 (C_{60}) is probably the most studied individual type of nanomaterial. The spherical shape of C_{60} is constructed from twelve pentagons and twenty hexagons and resembles a soccer ball ([\[link\]](#)a). The next stable higher fullerene is C_{70} ([\[link\]](#)b) that is shaped like a rugby or American football. The progression of higher fullerenes continues in the sequence C_{74} , C_{76} , C_{78} , etc. The structural relationship between each involves the addition of six membered rings. Mathematically (and chemically) two principles define the existence of a stable fullerene, i.e., Euler's theorem and isolated pentagon rule (IPR). Euler's theorem states that for the closure of each spherical network, n ($n \geq 2$) hexagons and 12 pentagons are required while the IPR says no two pentagons may be connected directly with each other as destabilization is caused by two adjacent pentagons.



(a)



(b)

Molecular structures of (a) C_{60} and (b) C_{70} .

Although fullerenes are composed of sp^2 carbons in a similar manner to graphite, fullerenes are soluble in various common organic solvents. Due to their hydrophobic nature, fullerenes are most soluble in CS_2 (C_{60} = 7.9 mg/mL) and toluene (C_{60} = 2.8 mg/mL). Although fullerenes have a conjugated system, their aromaticity is distinctive from benzene that has all C-C bonds of equal lengths, in fullerenes two distinct classes of bonds exist. The shorter bonds are at the junctions of two hexagons ([6, 6] bonds) and the longer bonds at the junctions of a hexagon and a pentagon ([5,6] bonds). This difference in bonding is responsible for some of the observed reactivity of fullerenes.

Synthesis of fullerenes

The first observation of fullerenes was in molecular beam experiments at Rice University. Subsequent studies demonstrated that C_{60} it was relatively easy to produce grams of fullerenes. Although the synthesis is relatively straightforward fullerene purification remains a challenge and determines fullerene's commercial price. The first method of production of measurable quantities of fullerenes used laser vaporization of carbon in an inert atmosphere, but this produced microscopic amounts of fullerenes. Laboratory scales of fullerene are prepared by the vaporization of carbon rods in a helium atmosphere. Commercial production ordinarily employs a simple ac or dc arc. The fullerenes in the black soot collected are extracted in toluene and purified by liquid chromatography. The magenta C_{60} comes off the column first, followed by the red C_{70} , and other higher fullerenes. Even though the mechanism of a carbon arc differs from that of a resistively heated carbon rod (because it involves a plasma) the He pressure for optimum C_{60} formation is very similar.

A ratio between the mass of fullerenes and the total mass of carbon soot defines fullerene yield. The yields determined by UV-Vis absorption are approximately 40%, 10-15%, and 15% in laser, electric arc, and solar processes. Interestingly, the laser ablation technique has both the highest yield and the lowest productivity and, therefore, a scale-up to a higher

power is costly. Thus, fullerene commercial production is a challenging task. The world's first computer controlled fullerene production plant is now operational at the MER Corporation, who pioneered the first commercial production of fullerene and fullerene products.

Endohedral fullerenes

Endohedral fullerenes are fullerenes that have incorporated in their inner sphere atoms, ions or clusters. Endohedral fullerenes are generally divided into two groups: endohedral metallofullerenes and non-metal doped fullerenes. The first endohedral metallofullerenes was called La@C_{60} . The @ sign in the name reflects the notion of a small molecule trapped inside a shell.

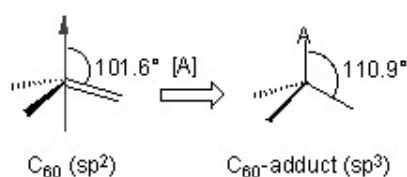
Doping fullerenes with metals takes place *in-situ* during the fullerene synthesis in an arc reactor or via laser evaporation. A wide range of metals have been encased inside a fullerene, i.e., Sc, Y, La, Ce, Ba, Sr, K, U, Zr, and Hf. Unfortunately, the synthesis of endohedral metallofullerenes is unspecific because in addition a high yield of unfilled fullerenes, compounds with different cage sizes are prepared (e.g., La@C_{60} or La@C_{82}). A characteristic of endohedral metallofullerenes is that electrons will transfer from the metal atom to the fullerene cage and that the metal atom takes a position off-center in the cage. The size of the charge transfer is not always simple to determine, but it is usually between 2 and 3 units (e.g., $\text{La}_2\text{@C}_{80}$) but can be as high as 6 electrons (e.g., $\text{Sc}_3\text{N@C}_{80}$). These anionic fullerene cages are very stable molecules and do not have the reactivity associated with ordinary empty fullerenes (see below). This lack of reactivity is utilized in a method to purify endohedral metallofullerenes from empty fullerenes.

The endohedral He@C_{60} and Ne@C_{60} form when C_{60} is exposed to a pressure of around 3 bar of the appropriate noble gases. Under these conditions it was possible to dope 1 in every 650,000 C_{60} cages with a helium atom. Endohedral complexes with He, Ne, Ar, Kr and Xe as well as numerous adducts of the He@C_{60} compound have also been proven with operating pressures of 3000 bars and incorporation of up to 0.1 % of the

noble gases. The isolation of N@C_{60} , N@C_{70} and P@C_{60} is very unusual and unlike the metal derivatives no charge transfer of the pnictide atom in the center to the carbon atoms of the cage takes place.

Chemically functionalized fullerenes

Although fullerenes have a conjugated aromatic system all the carbons are quaternary (i.e., containing no hydrogen), which results in making many of the characteristic substitution reactions of planar aromatics impossible. Thus, only two types of chemical transformations exist: redox reactions and addition reactions. Of these, addition reactions have the largest synthetic value. Another remarkable feature of fullerene addition chemistry is the thermodynamics of the process. Since the sp^2 carbon atoms in a fullerene are pyramidalized there is significant strain energy. For example, the strain energy in C_{60} is *ca* 8 kcal/mol, which is 80% of its heat of formation. So the relief of this strain energy leading to sp^3 hybridized C atoms is the major driving force for addition reactions ([\[link\]](#)). As a consequence, most additions to fullerenes are exothermic reactions.

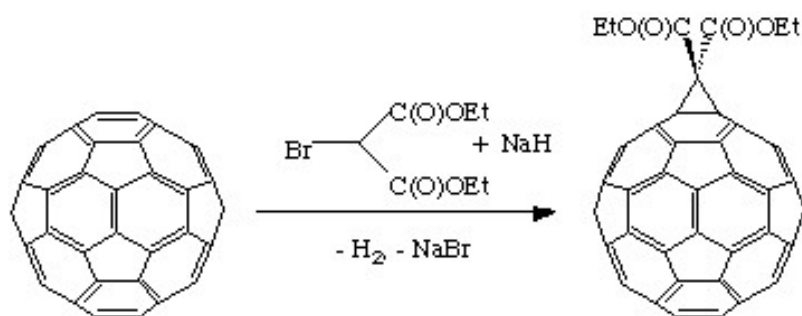


Strain release after
addition of reagent
A to a pyramidalize
carbon of C_{60} .

Cyclic voltammetry (CV) studies show that C_{60} can be reduced and oxidized reversibly up to 6 electrons with one-electron transfer processes. Fulleride anions can be generated by electrochemical method and then be

used to synthesize covalent organofullerene derivatives. Alkali metals can chemically reduce fullerene in solution and solid state to form M_xC_{60} ($x = 3 - 6$). C_{60} can also be reduced by less electropositive metals like mercury to form C_{60}^- and C_{60}^{2-} . In addition, salts can also be synthesized with organic molecules, for example $[TDAE^+][C_{60}^-]$ possesses interesting electronic and magnetic behavior.

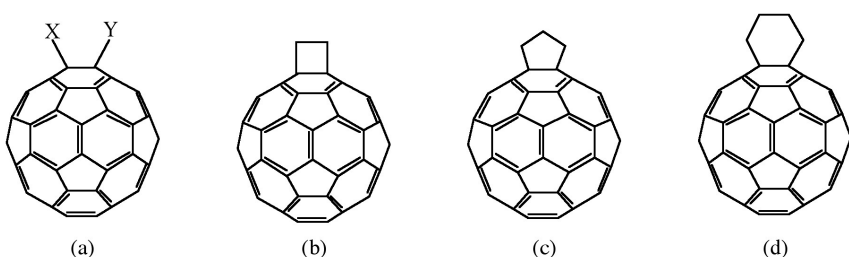
Geometric and electronic analysis predicted that fullerene behaves like an electro-poor conjugated polyolefin. Indeed C_{60} and C_{70} undergo a range of nucleophilic reactions with carbon, nitrogen, phosphorous and oxygen nucleophiles. C_{60} reacts readily with organolithium and Grignard compounds to form alkyl, phenyl or alkanyl fullerenes. Possibly the most widely used additions to fullerene is the Bingel reaction ([\[link\]](#)), where a carbon nucleophile, generated by deprotonation of α -halo malonate esters or ketones, is added to form a cyclopropanation product. The α -halo esters and ketones can also be generated in situ with I_2 or CBr_4 and a weak base as 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU). The Bingel reaction is considered one of the most versatile and efficient methods to functionalize C_{60} .



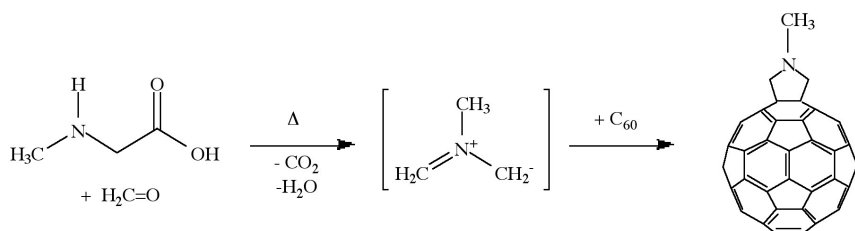
Bingel reaction of C_{60} with 2-bromoethylmalonate.

Cycloaddition is another powerful tool to functionalize fullerenes, in particular because of its selectivity with the 6,6 bonds, limiting the possible isomers ([\[link\]](#)). The dienophilic feature of the [6,6] double bonds of C_{60} enables the molecule to undergo various cycloaddition reactions in which

the monoadducts can be generated in high yields. The best studies cycloaddition reactions of fullerene are [3+2] additions with diazoderivatives and azomethine ylides (Prato reactions). In this reaction, azomethine ylides can be generated *in situ* from condensation of α -amino acids with aldehydes or ketones, which produce 1,3 dipoles to further react with C_{60} in good yields ([\[link\]](#)). Hundreds of useful building blocks have been generated by those two methods. The Prato reactions have also been successfully applied to carbon nanotubes.



Geometrical shapes built onto a [6,6] ring junction: a) open, b) four-membered ring, c) five-membered ring, and d) six-membered ring.



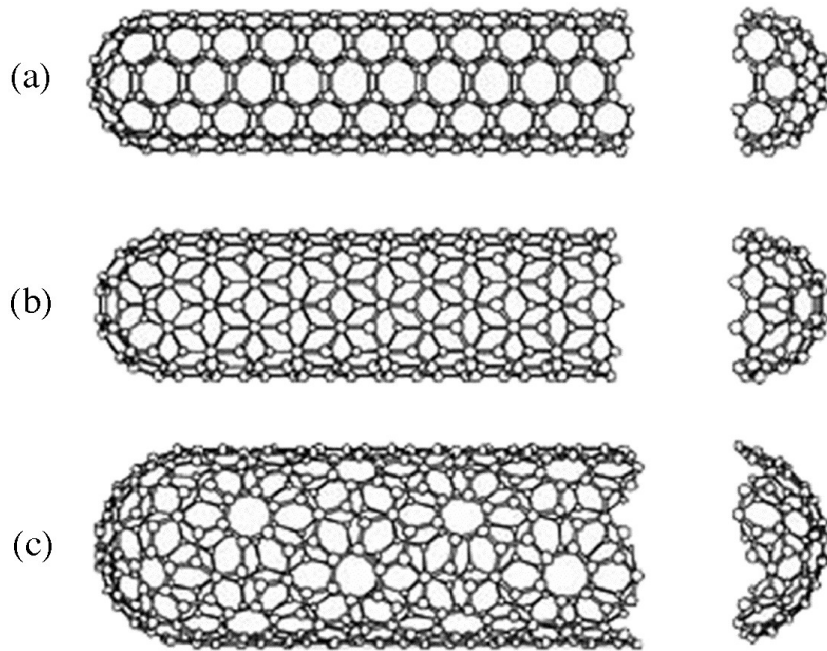
Prato reaction of C₆₀ with N-methylglycine and paraformaldehyde.

The oxidation of fullerenes, such as C_{60} , has been of increasing interest with regard to applications in photoelectric devices, biological systems, and possible remediation of fullerenes. The oxidation of C_{60} to $C_{60}O_n$ ($n = 1, 2$) may be accomplished by photooxidation, ozonolysis, and epoxidation. With each of these methods, there is a limit to the isolable oxygenated product, $C_{60}O_n$ with $n < 3$. Highly oxygenated fullerenes, $C_{60}O_n$ with $3 \leq n \leq 9$, have been prepared by the catalytic oxidation of C_{60} with $ReMeO_3/H_2O_2$.

Carbon nanotubes

A key breakthrough in carbon nanochemistry came in 1993 with the report of needle-like tubes made exclusively of carbon. This material became known as carbon nanotubes (CNTs). There are several types of nanotubes. The first discovery was of multi walled tubes (MWNTs) resembling many pipes nested within each other. Shortly after MWNTs were discovered single walled nanotubes (SWNTs) were observed. Single walled tubes resemble a single pipe that is potentially capped at each end. The properties of single walled and multi walled tubes are generally the same, although single walled tubes are believed to have superior mechanical strength and thermal and electrical conductivity; it is also more difficult to manufacture them.

Single walled carbon nanotubes (SWNTs) are by definition fullerene materials. Their structure consists of a graphene sheet rolled into a tube and capped by half a fullerene ([\[link\]](#)). The carbon atoms in a SWNT, like those in a fullerene, are sp^2 hybridized. The structure of a nanotube is analogous to taking this graphene sheet and rolling it into a seamless cylinder. The different types of SWNTs are defined by their diameter and chirality. Most of the presently used single-wall carbon nanotubes have been synthesized by the pulsed laser vaporization method, however, increasingly SWNTs are prepared by vapor liquid solid catalyzed growth.



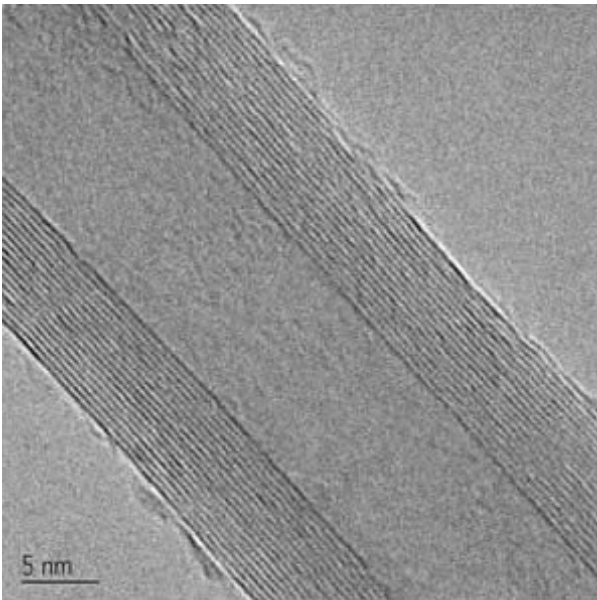
Structure of single walled carbon nanotubes (SWNTs) with (a) armchair, (b) zig-zag, and (c) chiral chirality.

The physical properties of SWNTs have made them an extremely attractive material for the manufacturing of nano devices. SWNTs have been shown to be stronger than steel as estimates for the Young's modulus approaches 1 Tpa. Their electrical conductance is comparable to copper with anticipate current densities of up to 10^{13} A/cm² and a resistivity as low as 0.34×10^{-4} Ω .cm at room temperatures. Finally, they have a high thermal conductivity (3000 - 6000 W.m/K).

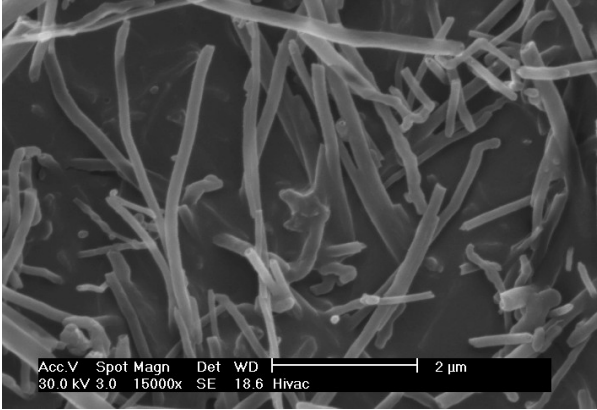
The electronic properties of a particular SWNT structure are based on its chirality or twist in the structure of the tube which is defined by its n,m value. The values of n and m determine the chirality, or "twist" of the nanotube. The chirality in turn affects the conductance of the nanotube, its density, its lattice structure, and other properties. A SWNT is considered metallic if the value $n-m$ is divisible by three. Otherwise, the nanotube is semi-conducting. The external environment also has an effect on the

conductance of a tube, thus molecules such as O_2 and NH_3 can change the overall conductance of a tube, while the presence of metals have been shown to significantly effect the opto-electronic properties of SWNTs.

Multi walled carbon nanotubes (MWNTs) range from double walled NTs, through many-walled NTs ([\[link\]](#)) to carbon nanofibers. Carbon nanofibers are the extreme of multi walled tubes ([\[link\]](#)) and they are thicker and longer than either SWNTs or MWNTs, having a cross-sectional of ca. 500 \AA^2 and are between 10 to 100 \mu m in length. They have been used extensively in the construction of high strength composites.



TEM image of an individual multi walled carbon nanotube (MWNTs). Copyright of Nanotech Innovations.



SEM image of vapor grown carbon nanofibers.

Synthesis of carbon nanotubes

A range of methodologies have been developed to produce nanotubes in sizeable quantities, including arc discharge, laser ablation, high pressure carbon monoxide (HiPco), and vapor liquid solid (VLS) growth. All these processes take place in vacuum or at low pressure with a process gases, although VLS growth can take place at atmospheric pressure. Large quantities of nanotubes can be synthesized by these methods; advances in catalysis and continuous growth processes are making SWNTs more commercially viable.

The first observation of nanotubes was in the carbon soot formed during the arc discharge production of fullerenes. The high temperatures caused by the discharge caused the carbon contained in the negative electrode to sublime and the CNTs are deposited on the opposing electrode. Tubes produced by this method were initially multi walled tubes (MWNTs). However, with the addition of cobalt to the vaporized carbon, it is possible to grow single walled nanotubes. This method it produces a mixture of components, and requires further purification to separate the CNTs from the soot and the residual catalytic metals. Producing CNTs in high yield depends on the

uniformity of the plasma arc, and the temperature of the deposit forming on the carbon electrode.

Higher yield and purity of SWNTs may be prepared by the use of a dual-pulsed laser. SWNTs can be grown in a 50% yield through direct vaporization of a Co/Ni doped graphite rod with a high-powered laser in a tube furnace operating at 1200 °C. The material produced by this method appears as a mat of “ropes”, 10 - 20 nm in diameter and up to 100 μm or more in length. Each rope consists of a bundle of SWNTs, aligned along a common axis. By varying the process parameters such as catalyst composition and the growth temperature, the average nanotube diameter and size distribution can be varied. Although arc-discharge and laser vaporization are currently the principal methods for obtaining small quantities of high quality SWNTs, both methods suffer from drawbacks. The first is that they involve evaporating the carbon source, making scale-up on an industrial level difficult and energetically expensive. The second issue relates to the fact that vaporization methods grow SWNTs in highly tangled forms, mixed with unwanted forms of carbon and/or metal species. The SWNTs thus produced are difficult to purify, manipulate, and assemble for building nanotube-device architectures for practical applications.

In order to overcome some of the difficulties of these high-energy processes, the chemical catalysis method was developed in which a hydrocarbon feedstock is used in combination with a metal catalyst. The catalyst is typically, but not limited to iron, cobalt, or iron/molybdenum, it is heated under reducing conditions in the presence of a suitable carbon feedstock, e.g., ethylene. This method can be used for both SWNTs and MWNTs; the formation of each is controlled by the identity of the catalyst and the reaction conditions. A convenient laboratory scale apparatus is available from Nanotech Innovations, Inc., for the synthesis of highly uniform, consistent, research sample that uses pre-weighed catalyst/carbon source ampoules. This system, allows for 200 mg samples of MWNTs to be prepared for research and testing. The use of CO as a feedstock, in place of a hydrocarbon, led to the development of the high-pressure carbon monoxide (HiPco) procedure for SWNT synthesis. By this method, it is possible to produce gram quantities of SWNTs, unfortunately, efforts to scale beyond that have not met with complete success.

Initially developed for small-scale investigations of catalyst activity, vapor liquid solid (VLS) growth of nanotubes has been highly studied, and now shows promise for large-scale production of nanotubes. Recent approaches have involved the use of well-defined nanoparticle or molecular precursors and many different transition metals have been employed, but iron, nickel, and cobalt remain to be the focus of most research. The nanotubes grow at the sites of the metal catalyst; the carbon-containing gas is broken apart at the surface of the catalyst particle, and the carbon is transported to the edges of the particle, where it forms the nanotube. The length of the tube grown in surface supported catalyst VLS systems appears to be dependent on the orientation of the growing tube with the surface. By properly adjusting the surface concentration and aggregation of the catalyst particles it is possible to synthesize vertically aligned carbon nanotubes, i.e., as a carpet perpendicular to the substrate.

Of the various means for nanotube synthesis, the chemical processes show the greatest promise for industrial scale deposition in terms of its price/unit ratio. There are additional advantages to the VLS growth, which unlike the other methods is capable of growing nanotubes directly on a desired substrate. The growth sites are controllable by careful deposition of the catalyst. Additionally, no other growth methods have been developed to produce vertically aligned SWNTs.

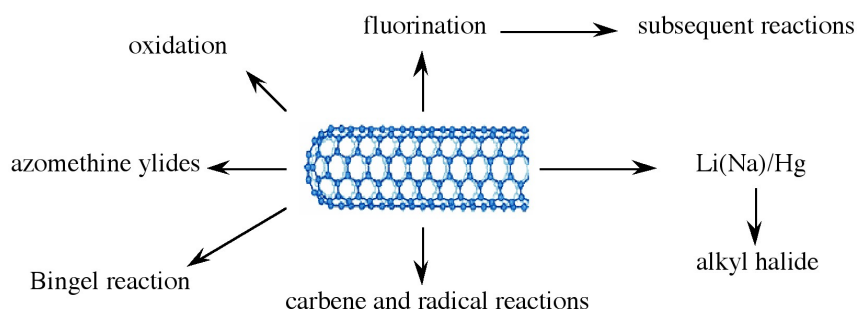
Chemical functionalization of carbon nanotubes

The limitation of using carbon nanotubes in any practical applications has been its solubility; for example SWNTs have little to no solubility in most solvent due to the aggregation of the tubes. Aggregation/roping of nanotubes occurs as a result of the high van der Waals binding energy of *ca.* 500 eV per mm of tube contact. The van der Waals force between the tubes is so great, that it take tremendous energy to pry them apart, making it very to make combination of nanotubes with other materials such as in composite applications. The functionalization of nanotubes, i.e., the attachment of “chemical functional groups” provides the path to overcome these barriers. Functionalization can improve solubility as well as processibility, and has been used to align the properties of nanotubes to

those of other materials. The clearest example of this is the ability to solubilize nanotubes in a variety of solvents, including water. It is important when discussing functionalization that a distinction is made between covalent and non-covalent functionalization.

Current methods for solubilizing nanotubes without covalent functionalization include highly aromatic solvents, super acids, polymers, or surfactants. Non-covalent “functionalization” is generally on the concept of supramolecular interactions between the SWNT and some macromolecule as a result of various adsorption forces, such as van der Waals’ and π -stacking interactions. The chemical speciation of the nanotube itself is not altered as a result of the interaction. In contrast, covalent functionalization relies on the chemical reaction at either the sidewall or end of the SWNT. As may be expected the high aspect ratio of nanotubes means that sidewall functionalization is much more important than the functionalization of the cap. Direct covalent sidewall functionalization is associated with a change of hybridization from sp^2 to sp^3 and a simultaneous loss of conjugation. An alternative approach to covalent functionalization involves the reaction of defects present (or generated) in the structure of the nanotube. Defect sites can be the open ends and holes in the sidewalls, and pentagon and heptagon irregularities in the hexagon graphene framework (often associated with bends in the tubes). All these functionalizations are exohedral derivatizations. Taking the hollow structure of nanotubes into consideration, endohedral functionalization of SWNTs is possible, i.e., the filling of the tubes with atoms or small molecules. It is important to note that covalent functionalization methods have one problem in common: extensive covalent functionalization modifies SWNT properties by disrupting the continuous π -system of SWNTs.

Various applications of nanotubes require different, specific modification to achieve desirable physical and chemical properties of nanotubes. In this regard, covalent functionalization provides a higher degree of fine-tuning the chemistry and physics of SWNTs than non-covalent functionalization. Until now, a variety of methods have been used to achieve the functionalization of nanotubes ([\[link\]](#)).



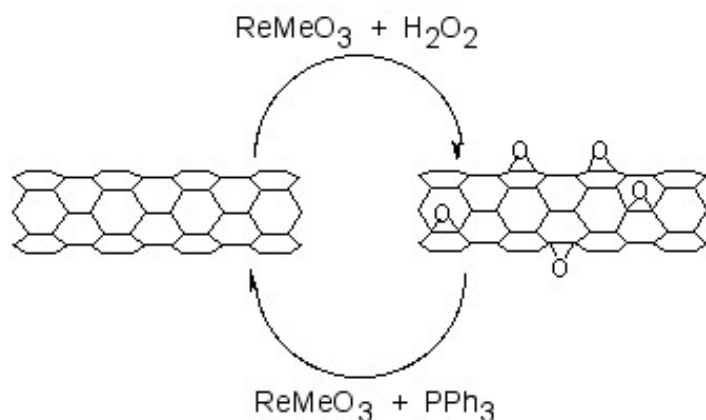
Schematic description of various covalent functionalization strategies for SWNTs.

Taking chemistry developed for C_{60} , SWNTs may be functionalized using 1,3 dipolar addition of azomethine ylides. The functionalized SWNTs are soluble in most common organic solvents. The azomethine ylide functionalization method was also used for the purification of SWNTs. Under electrochemical conditions, aryl diazonium salts react with SWNTs to achieve functionalized SWNTs, alternatively the diazonium ions may be generated *in-situ* from the corresponding aniline, while a solvent free reaction provides the best chance for large-scale functionalization this way. In each of these methods it is possible to control the amount of functionalization on the tube by varying reaction times and the reagents used; functionalization as high as 1 group per every 10 - 25 carbon atoms is possible.

Organic functionalization through the use of alkyl halides, a radical pathway, on tubes treated with lithium in liquid ammonia offers a simple and flexible route to a range of functional groups. In this reaction, functionalization occurs on every 17 carbons. Most success has been found when the tubes are dodecylated. These tubes are soluble in chloroform, DMF, and THF.

The addition of oxygen moieties to SWNT sidewalls can be achieved by treatment with acid or wet air oxidation, and ozonolysis. The direct epoxidation of SWNTs may be accomplished by the direct reaction with a peroxide reagent, or catalytically. Catalytic de-epoxidation ([\[link\]](#)) allows

for the quantitative analysis of sidewall epoxide and led to the surprising result that previously assumed “pure” SWNTs actually contain *ca.* 1 oxygen per 250 carbon atoms.



Catalytic oxidation and de-epoxidation of SWNTs.

One of the easiest functionalization routes, and a useful synthon for subsequent conversions, is the fluorination of SWNTs, using elemental fluorine. Importantly, a C:F ratios of up to 2:1 can be achieved without disruption of the tubular structure. The fluorinated SWNTs (F-SWNTs) proved to be much more soluble than pristine SWNTs in alcohols (1 mg/mL in *iso*-propanol), DMF and other selected organic solvents. Scanning tunneling microscopy (STM) revealed that the fluorine formed bands of approximately 20 nm, while calculations using DFT revealed 1,2 addition is more energetically preferable than 1,4 addition, which has been confirmed by solid state ^{13}C NMR. F-SWNTs make highly flexible synthons and subsequent elaboration has been performed with organo lithium, Grignard reagents, and amines.

Functionalized nanotubes can be characterized by a variety of techniques, such as atomic force microscopy (AFM), transmission electron microscopy (TEM), UV-vis spectroscopy, and Raman spectroscopy. Changes in the

Raman spectrum of a nanotube sample can indicate if functionalization has occurred. Pristine tubes exhibit two distinct bands. They are the radial breathing mode (230 cm^{-1}) and the tangential mode (1590 cm^{-1}). When functionalized, a new band, called the disorder band, appears at *ca.* 1350 cm^{-1} . This band is attributed to sp^3 -hybridized carbons in the tube. Unfortunately, while the presence of a significant D mode is consistent with sidewall functionalization and the relative intensity of D (disorder) mode versus the tangential G mode ($1550 - 1600\text{ cm}^{-1}$) is often used as a measure of the level of substitution. However, it has been shown that Raman is an unreliable method for determination of the extent of functionalization since the relative intensity of the D band is also a function of the substituents distribution as well as concentration. Recent studies suggest that solid state ^{13}C NMR are possibly the only definitive method of demonstrating covalent attachment of particular functional groups.

Coating carbon nanotubes: creating inorganic nanostructures

Fullerenes, nanotubes and nanofibers represent suitable substrates for the seeding other materials such as oxides and other minerals, as well as semiconductors. In this regard, the carbon nanomaterial acts as a seed point for the growth as well as a method of defining unusual aspect ratios. For example, silica fibers can be prepared by a number of methods, but it is only through coating SWNTs that silica nano-fibers with of micron lengths with tens of nanometers in diameter may be prepared.

While C_{60} itself does not readily seed the growth of inorganic materials, liquid phase deposition of oxides, such as silica, in the presence of fulleranol, $\text{C}_{60}(\text{OH})_n$, results in the formation of uniform oxide spheres. It appears the fulleranol acts as both a reagent and a physical point for subsequent oxide growth, and it is C_{60} , or an aggregate of C_{60} , that is present within the spherical particle. The addition of fulleranol alters the morphology and crystal phase of CaCO_3 precipitates from aqueous solution, resulting in the formation of spherical features, 5-pointed flower shaped clusters, and triangular crystals as opposed to the usual rhombic crystals. In addition, the meta-stable vaterite phase is observed with the addition of $\text{C}_{60}(\text{OH})_n$.

As noted above individual SWNTs may be obtained in solution when encased in a cylindrical micelle of a suitable surfactant. These individualized nanotubes can be coated with a range of inorganic materials. Liquid phase deposition (LPD) appears to have significant advantages over other methods such as incorporating surfacted SWNTs into a preceramic matrix, *in situ* growth of the SWNT in an oxide matrix, and sol-gel methods. The primary advantage of LPD growth is that individual SWNTs may be coated rather than bundles or ropes. For example, SWNTs have been coated with silica by liquid phase deposition (LPD) using a silica/H₂SiF₆ solution and a surfactant-stabilized solution of SWNTs. The thickness of the coating is dependent on the reaction mixture concentration and the reaction time. The SWNT core can be removed by thermolysis under oxidizing conditions to leave a silica nano fiber. It is interesting to note that the use of a surfactant is counter productive when using MWNTs and VGFs, in this case surface activation of the nanotube offers the suitable growth initiation. Pre-oxidation of the MWNT or VGF allows for uniform coatings to be deposited. The coated SWNTs, MWNTs, and VGFs can be subsequently reacted with suitable surface reagents to impart miscibility in aqueous solutions, guar gels, and organic matrixes. In addition to simple oxides, coated nanotubes have been prepared with minerals such as carbonates and semiconductors.

Bibliography

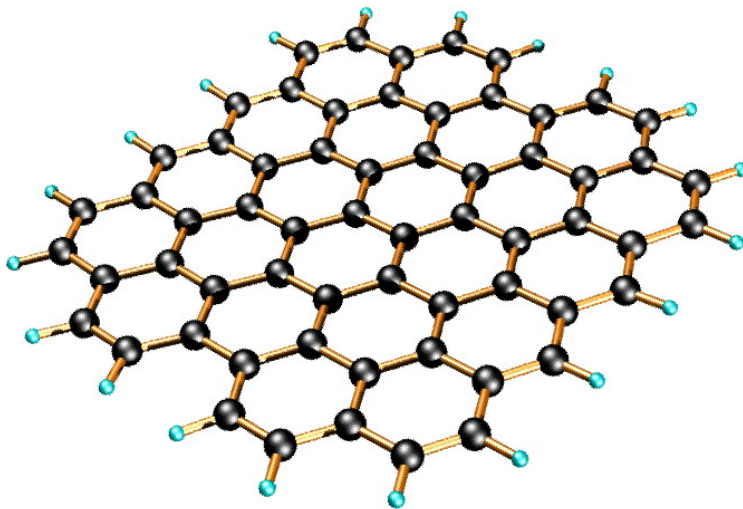
- S. M. Bachilo, M. S. Strano, C. Kittrell, R. H. Hauge, R. E. Smalley, and R. B. Weisman, *Science*, 2002, **298**, 2361.
- D. S. Bethune, C. H. Klang, M. S. deVries, G. Gorman, R. Savoy, J. Vazquez, and R. Beyers, *Nature*, 1993, **363**, 605.
- J. J. Brege, C. Gallaway, and A. R. Barron, *J. Phys. Chem., C*, 2007, **111**, 17812.
- C. A. Dyke and J. M. Tour, *J. Am. Chem. Soc.*, 2003, **125**, 1156.
- Z. Ge, J. C. Duchamp, T. Cai, H. W. Gibson, and H. C. Dorn, *J. Am. Chem. Soc.*, 2005, **127**, 16292.
- L. A. Girifalco, M. Hodak, and R. S. Lee, *Phys. Rev. B*, 2000, **62**, 13104.
- T. Guo, P. Nikolaev, A. G. Rinzler, D. Tománek, D. T. Colbert, and R. E. Smalley, *J. Phys. Chem.*, 1995, **99**, 10694.

- J. H. Hafner, M. J. Bronikowski, B. R. Azamian, P. Nikolaev, A. G. Rinzler, D. T. Colbert, K. A. Smith, and R. E. Smalley, *Chem. Phys. Lett.*, 1998, **296**, 195.
- A. Hirsch, *Angew. Chem. Int. Ed.*, 2002, **40**, 4002.
- S. Iijima and T. Ichihashi, *Nature*, 1993, **363**, 603.
- H. R. Jafry, E. A. Whitsitt, and A. R. Barron, *J. Mater. Sci.*, 2007, **42**, 7381.
- H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley, *Nature*, 1985, **318**, 162.
- F. Liang, A. K. Sadana, A. Peera, J. Chattopadhyay, Z. Gu, R. H. Hauge, and W. E. Billups, *Nano Lett.*, 2004, **4**, 1257.
- D. Ogrin and A. R. Barron, *J. Mol. Cat. A: Chem.*, 2006, **244**, 267.
- D. Ogrin, J. Chattopadhyay, A. K. Sadana, E. Billups, and A. R. Barron, *J. Am. Chem. Soc.*, 2006, **128**, 11322.
- R. E. Smalley, *Acc. Chem. Res.*, 1992, **25**, 98.
- M. M. J. Treacy, T. W. Ebbesen, and J. M. Gibson, *Nature*, 1996, **381**, 678.
- E. A. Whitsitt and A. R. Barron, *Nano Lett.*, 2003, **3**, 775.
- J. Yang and A. R. Barron, *Chem. Commun.*, 2004, 2884.
- L. Zeng, L. B. Alemany, C. L. Edwards, and A. R. Barron, *Nano Res.*, 2008, **1**, 72.

Graphene

Introduction

Graphene is a one-atom-thick planar sheet of sp^2 -bonded carbon atoms that are densely packed in a honeycomb crystal lattice ([\[link\]](#)). The name comes from “graphite” and “alkene”; graphite itself consists of many graphene sheets stacked together.



Idealized structure of a single graphene sheet.

Single-layer graphene nanosheets were first characterized in 2004, prepared by mechanical exfoliation (the “scotch-tape” method) of bulk graphite. Later graphene was produced by epitaxial chemical vapor deposition on silicon carbide and nickel substrates. Most recently, graphene nanoribbons (GNRs) have been prepared by the oxidative treatment of carbon nanotubes and by plasma etching of nanotubes embedded in polymer films.

Physical properties of graphene

Graphene has been reported to have a Young's modulus of 1 TPa and intrinsic strength of 130 GP; similar to single walled carbon nanotubes (SWNTs). The electronic properties of graphene also have some similarity with carbon nanotubes. Graphene is a zero-bandgap semiconductor. Electron mobility in graphene is extraordinarily high ($15,000 \text{ cm}^2/\text{V.s}$ at room temperature) and ballistic electron transport is reported to be on length scales comparable to that of SWNTs. One of the most promising aspects of graphene involves the use of GNRs. Cutting an individual graphene layer into a long strip can yield semiconducting materials where the bandgap is tuned by the width of the ribbon.

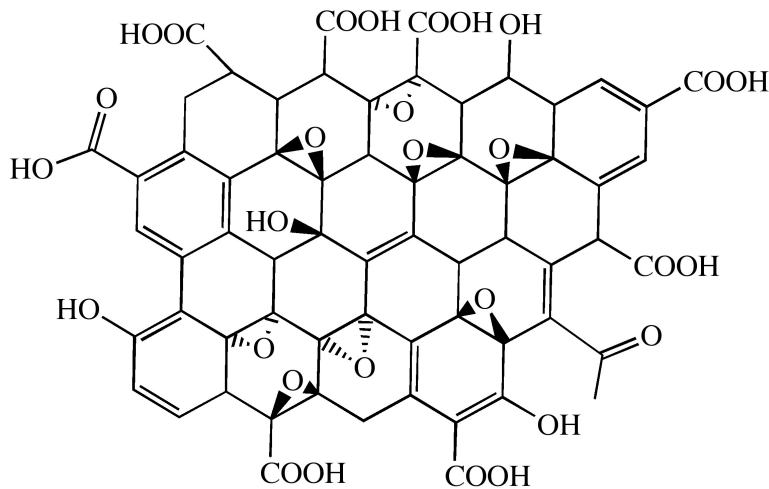
While graphene's novel electronic and physical properties guarantee this material will be studied for years to come, there are some fundamental obstacles yet to overcome before graphene based materials can be fully utilized. The aforementioned methods of graphene preparation are effective; however, they are impractical for large-scale manufacturing. The most plentiful and inexpensive source of graphene is bulk graphite. Chemical methods for exfoliation of graphene from graphite provide the most realistic and scalable approach to graphene materials.

Graphene layers are held together in graphite by enormous van der Waals forces. Overcoming these forces is the major obstacle to graphite exfoliation. To date, chemical efforts at graphite exfoliation have been focused primarily on intercalation, chemical derivatization, thermal expansion, oxidation-reduction, the use of surfactants, or some combination of these.

Graphite oxide

Probably the most common route to graphene involves the production of graphite oxide (GO) by extremely harsh oxidation chemistry. The methods of Staudenmeier or Hummers are most commonly used to produce GO, a highly exfoliated material that is dispersible in water. The structure of GO has been the subject of numerous studies; it is known to contain epoxide functional groups along the basal plane of sheets as well as hydroxyl and carboxyl moieties along the edges ([\[link\]](#)). In contrast to other methods for the synthesis of GO, the the *m*-peroxybenzoic acid (*m*-CPBA) oxidation of

microcrystalline synthetic graphite at room temperature yields graphite epoxide in high yield, without significant additional defects.



Idealized structure proposed for graphene oxide (GO). Adapted from C. E. Hamilton, PhD Thesis, Rice University (2009).

As graphite oxide is electrically insulating, it must be converted by chemical reduction to restore the electronic properties of graphene. Chemically converted graphene (CCG) is typically reduced by hydrazine or borohydride. The properties of CCG can never fully match those of graphene for two reasons:

1. Oxidation to GO introduces defects.
2. Chemical reduction does not fully restore the graphitic structure.

As would be expected, CCG is prone to aggregation unless stabilized. Graphene materials produced from pristine graphite avoid harsh oxidation to GO and subsequent (incomplete) reduction; thus, materials produced are potentially much better suited to electronics applications.

A catalytic approach to the removal of epoxides from fullerenes and SWNTs has been applied to graphene epoxide and GO. Treatment of oxidized graphenes with methyltrioxorhenium (MeReO_3 , MTO) in the presence of PPh_3 results in the oxygen transfer, to form O=PPh_3 and allow for quantification of the C:O ratio.

Homogeneous graphene dispersions

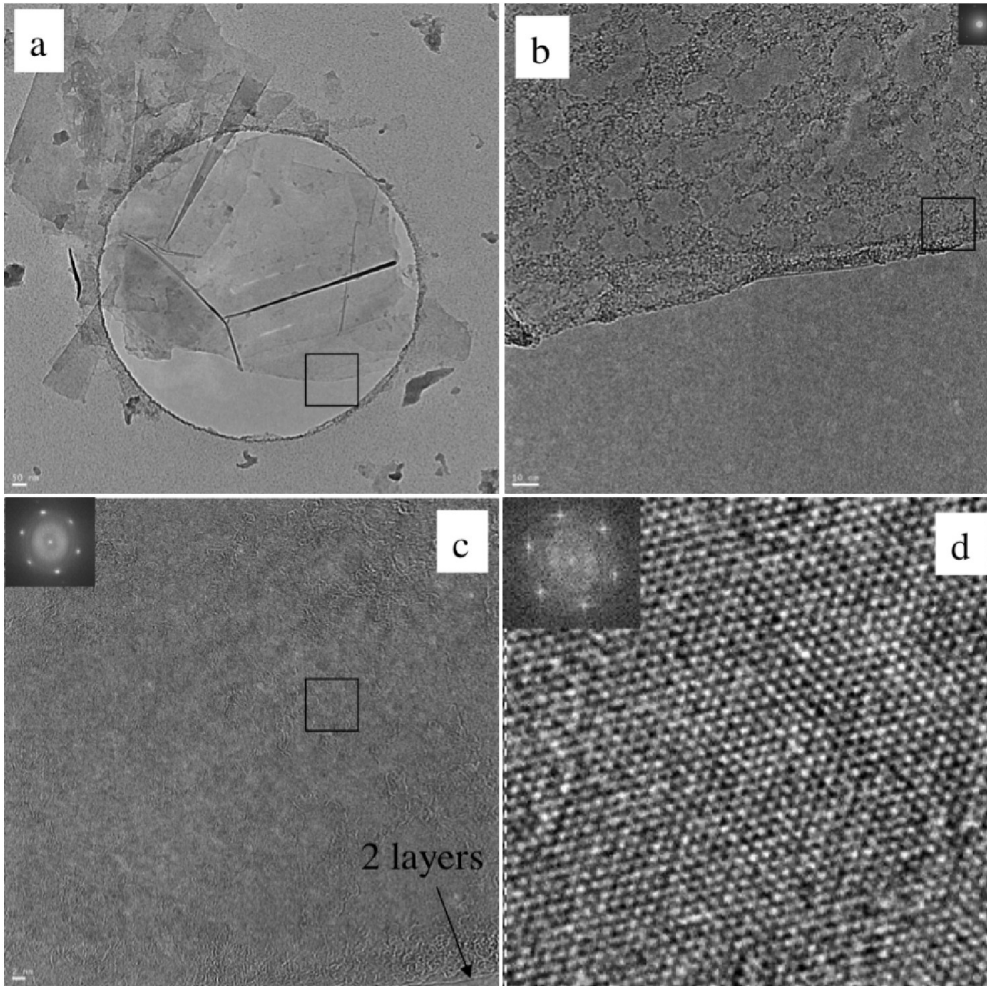
An alternate approach to producing graphene materials involves the use of pristine graphite as starting material. The fundamental value of such an approach lies in its avoidance of oxidation to GO and subsequent (incomplete) reduction, thereby preserving the desirable electronic properties of graphene. There is precedent for exfoliation of pristine graphite in neat organic solvents without oxidation or surfactants. It has been reported that *N,N*-dimethylformamide (DMF) dispersions of graphene are possible, but no detailed characterization of the dispersions were reported. In contrast, Coleman and coworkers reported similar dispersions using *N*-methylpyrrolidone (NMP), resulting in individual sheets of graphene at a concentration of ≤ 0.01 mg/mL. NMP and DMF are highly polar solvents, and not ideal in cases where reaction chemistry requires a nonpolar medium. Further, they are hygroscopic, making their use problematic when water must be excluded from reaction mixtures. Finally, DMF is prone to thermal and chemical decomposition.

Recently, dispersions of graphene has been reported in *ortho*-dichlorobenzene (ODCB) using a wide range of graphite sources. The choice of ODCB for graphite exfoliation was based on several criteria:

1. ODCB is a common reaction solvent for fullerenes and is known to form stable SWNT dispersions.
2. ODCB is a convenient high-boiling aromatic, and is compatible with a variety of reaction chemistries.
3. ODCB, being aromatic, is able to interact with graphene *via* π - π stacking.
4. It has been suggested that good solvents for graphite exfoliation should have surface tension values of 40 – 50 mJ/m². ODCB has a surface tension of 36.6 mJ/m², close to the proposed range.

Graphite is readily exfoliated in ODCB with homogenization and sonication. Three starting materials were successfully dispersed: microcrystalline synthetic, thermally expanded, and highly ordered pyrolytic graphite (HOPG). Dispersions of microcrystalline synthetic graphite have a concentration of 0.03 mg/mL, determined gravimetrically. Dispersions from expanded graphite and HOPG are less concentrated (0.02 mg/mL).

High resolution transmission electron microscopy (HRTEM) shows mostly few-layer graphene ($n < 5$) with single layers and small flakes stacked on top ([\[link\]](#)). Large graphitic domains are visible; this is further supported by selected area electron diffraction (SAED) and fast Fourier transform (FFT) in selected areas. Atomic force microscope (AFM) images of dispersions sprayed onto silicon substrates shows extremely thin flakes with nearly all below 10 nm. Average height is 7 - 10 nm. The thinnest are less than 1 nm, graphene monolayers. Lateral dimensions of nanosheets range from 100 – 500 nm.



TEM images of single layer graphene from HOPG dispersion. (a) monolayer and few layer of graphene stacked with smaller flakes; (b) selected edge region from (a), (c) selected area from (b) with FFT inset, (d) HRTEM of boxed region in (c) showing lattice fringes with FFT inset. Adapted from C. E. Hamilton, PhD Thesis, Rice University (2009).

As-deposited films cast from ODCB graphene show poor electrical conductivity, however, after vacuum annealing at 400 °C for 12 hours the films improve vastly, having sheet resistances on the order of 60 Ω/sq . By

comparison, graphene epitaxially grown on Ni has a reported sheet resistance of 280 Ω/sq .

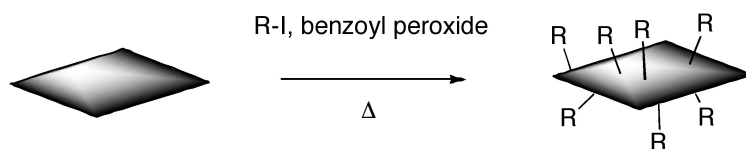
Covalent functionalization of graphene and graphite oxide

The covalent functionalization of SWNTs is well established. Some routes to covalently functionalized SWNTs include esterification/ amidation, reductive alkylation (Billups reaction), and treatment with azomethine ylides (Prato reaction), diazonium salts, or nitrenes. Conversely, the chemical derivatization of graphene and GO is still relatively unexplored.

Some methods previously demonstrated for SWNTs have been adapted to GO or graphene. GO carboxylic acid groups have been converted into acyl chlorides followed by amidation with long-chain amines. Additionally, the coupling of primary amines and amino acids via nucleophilic attack of GO epoxide groups has been reported. Yet another route coupled isocyanates to carboxylic acid groups of GO. Functionalization of partially reduced GO by aryldiazonium salts has also been demonstrated. The Billups reaction has been performed on the intercalation compound potassium graphite (C_8K), as well as graphite fluoride, and most recently GO. Graphene alkylation has been accomplished by treating graphite fluoride with alkyllithium reagents.

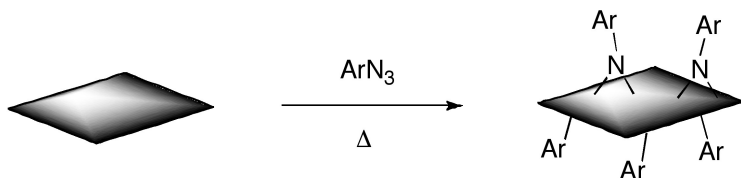
ODCB dispersions of graphene may be readily converted to covalently functionalize graphene. Thermal decomposition of benzoyl peroxide is used to initiate radical addition of alkyl iodides to graphene in ODCB dispersions.

Equation:



Additionally, functionalized graphene with nitrenes generated by thermal decomposition of aryl azides

Equation:



Bibliography

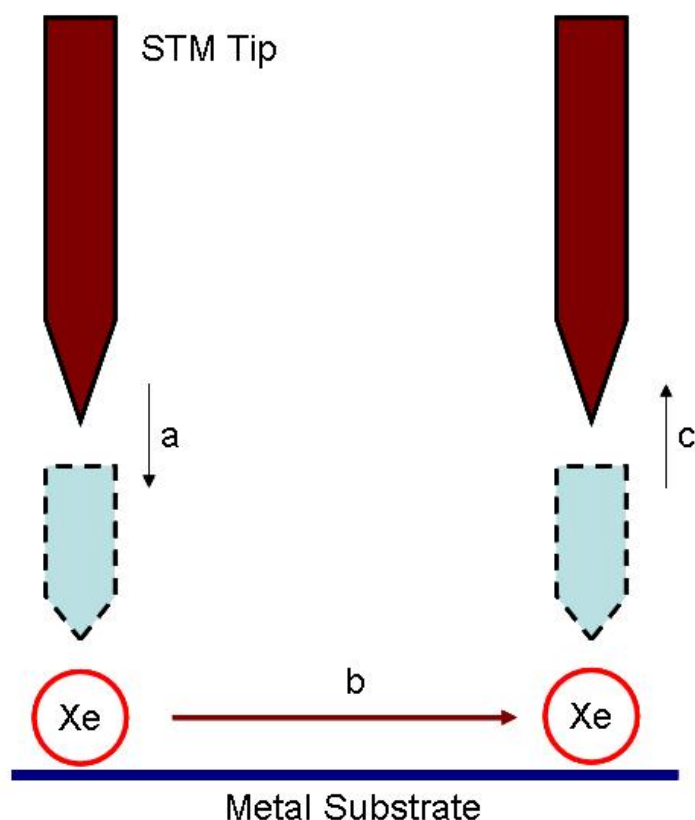
- P. Blake, P. D. Brimicombe, R. R. Nair, T. J. Booth, D. Jiang, F. Schedin, L. A. Ponomarenko, S. V. Morozov, H. F. Gleeson, E. W. Hill, A. K. Geim, and K. S. Novoselov, *Nano Lett.*, 2008, **8**, 1704.
- J. Chattopadhyay, A. Mukherjee, C. E. Hamilton, J.-H. Kang, S. Chakraborty, W. Guo, K. F. Kelly, A. R. Barron, and W. E. Billups, *J. Am. Chem. Soc.*, 2008, **130**, 5414.
- G. Eda, G. Fanchini, and M. Chhowalla, *Nat. Nanotechnol.*, 2008, **3**, 270.
- M. Y. Han, B. Ozyilmaz, Y. Zhang, and P. Kim, *Phys. Rev. Lett.*, 2008, **98**, 206805.
- Y. Hernandez, V. Nicolosi, M. Lotya, F. M. Blighe, Z. Sun, S. De, I. T. McGovern, B. Holland, M. Byrne, Y. K. Gun'Ko, J. J. Boland, P. Niraj, G. Duesberg, S. Krishnamurthy, R. Goodhue, J. Hutchinson, V. Scardaci, A. C. Ferrari, and J. N. Coleman, *Nat. Nanotechnol.*, 2008, **3**, 563.
- W. S. Hummers and R. E. Offeman, *J. Am. Chem. Soc.*, 1958, **80**, 1339.
- L. Jiao, L. Zhang, X. Wang, G. Diankov, and H. Dai, *Nature*, 2009, **458**, 877.
- D. V. Kosynkin, A. L. Higginbotham, A. Sinitskii, J. R. Lomeda, A. Dimiev, B. K. Price, and J. M. Tour, *Nature*, 2009, **458**, 872.
- D. Li, M. B. Mueller, S. Gilje, R. B. Kaner, and G. G. Wallace, *Nat. Nanotechnol.*, 2008, **3**, 101.
- S. Niyogi, E. Bekyarova, M. E. Itkis, J. L. McWilliams, M. A. Hamon, and R. C. Haddon, *J. Am. Chem. Soc.*, 2006, **128**, 7720.
- Y. Si and E. T. Samulski, *Nano Lett.*, 2008, **8**, 1679.
- L. Staudenmaier, *Ber. Dtsch. Chem. Ges.*, 1898, **31**, 1481.

Rolling Molecules on Surfaces Under STM Imaging

Introduction to surface motions at the molecular level

As single molecule imaging methods such as scanning tunneling microscope (STM), atomic force microscope (AFM), and transmission electron microscope (TEM) developed in the past decades, scientists have gained powerful tools to explore molecular structures and behaviors in previously unknown areas. Among these imaging methods, STM is probably the most suitable one to observe detail at molecular level. STM can operate in a wide range of conditions, provides very high resolution, and able to manipulate molecular motions with the tip. An interesting early example came from IBM in 1990, in which the STM was used to position individual atoms for the first time, spelling out "I-B-M" in Xenon atoms. This work revealed that observation and control of single atoms and molecular motions on surfaces were possible.

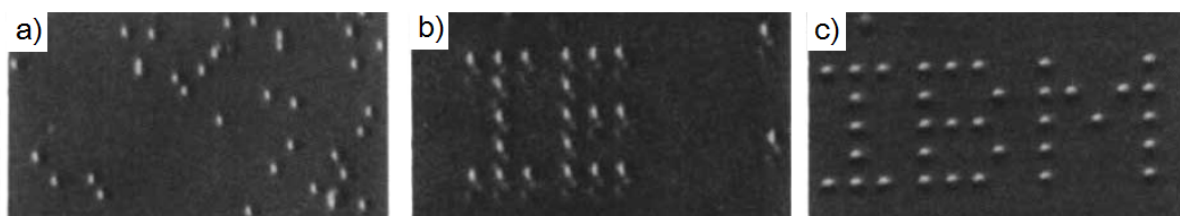
The IBM work, and subsequent experiments, relied on the fact that STM tip always exerts a finite force toward an adsorbate atom that contains both van der Waals and electrostatic forces was utilized for manipulation purpose. By adjusting the position and the voltage of the tip, the interactions between the tip and the target molecule were changed. Therefore, applying/releasing force to a single atom and make it move was possible [\[link\]](#).



Manipulation of STM tip toward a xenon atom. a) STM tip move onto a target atom then change the voltage and current of the tip to apply a stronger interaction. b) Move the atom to a desire position. c) After reaching the desire position, the tip released by switching back to the scanning voltage and current.

The actual positioning experiment was carried out in the following process. The nickel metal substrate was prepared by cycles of argon-ion sputtering, followed by annealing in a partial pressure of oxygen to remove surface carbon and other impurities. After the cleaning process, the sample was cooled to 4 K, and imaged with the STM to ensure the quality of surface.

The nickel sample was then doped with xenon. An image of the doped sample was taken at constant-current scanning conditions. Each xenon atom appears as a located randomly 1.6 Å high bump on the surface ([link](#)a). Under the imaging conditions (tip bias = 0.010 V with tunneling current 10^{-9} A) the interaction of the xenon with the tip is too weak to cause the position of the xenon atom to be perturbed. To move an atom, the STM tip was placed on top of the atom performing the procedure depicted in [link](#) to move it to its target. Repeating this process again and again led the researcher to build of the structure they desired [link](#)b and c.



Manipulation of STM tip starting with a) randomly dosed xenon sample, b) under construction - move xenon atom to desire position, and c) accomplishment of the manipulation. Adapted from D. M. Eigler and E. K. Schweizer, *Nature*, 1990, **344**, 524.

All motions on surfaces at the single molecule level can be described as by the following (or combination of the following) modes:

- i. Sliding.
- ii. Hopping.
- iii. Rolling.
- iv. Pivoting.

Although the power of STM imaging has been demonstrated, imaging of molecules themselves is still often a difficult task. The successful imaging of the IBM work was attributed to selection of a heavy atom. Other synthetic organic molecules without heavy atoms are much more difficult to be imaged under STM. Determinations of the mechanism of molecular

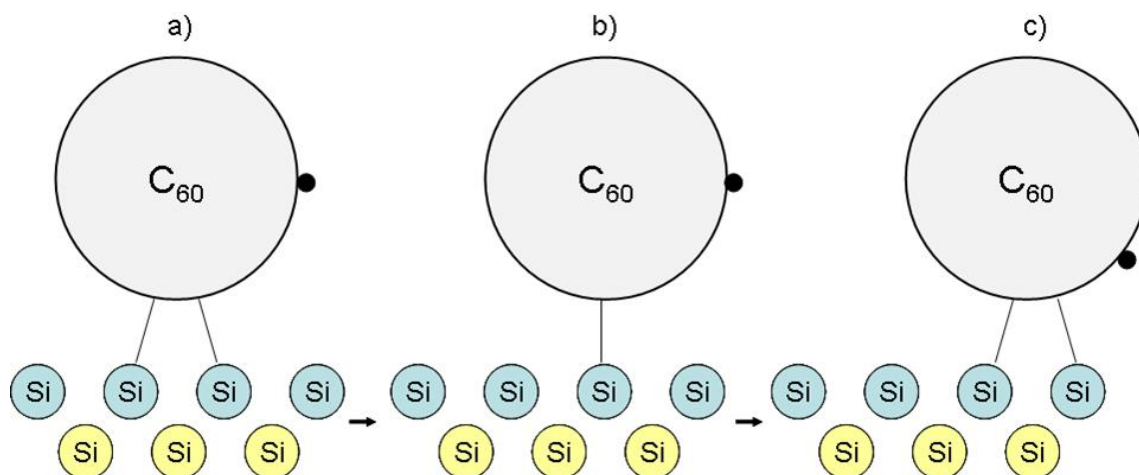
motion is another. Besides imaging methods themselves, other auxiliary methods such as DFT calculations and imaging of properly designed molecules are required to determine the mechanism by which a particular molecule moves across a surface.

Herein, we are particularly interested in surface-rolling molecules, i.e., those that are designed to roll on a surface. It is straightforward to imagine that if we want to construct (and image) surface-rolling molecules, we must think of making highly symmetrical structures. In addition, the magnitudes of interactions between the molecules and the surfaces have to be adequate; otherwise the molecules will be more susceptible to slide/hop or stick on the surfaces, instead of rolling. As a result, only very few molecules are known can roll and be detected on surfaces.

Surface rolling of molecules under the manipulation of STM tips

As described above, rolling motions are most likely to be observed on molecules having high degree of symmetry and suitable interactions between themselves and the surface. C_{60} is not only a highly symmetrical molecule but also readily imageable under STM due to its size. These properties together make C_{60} and its derivatives highly suitable to study with regards to surface-rolling motion.

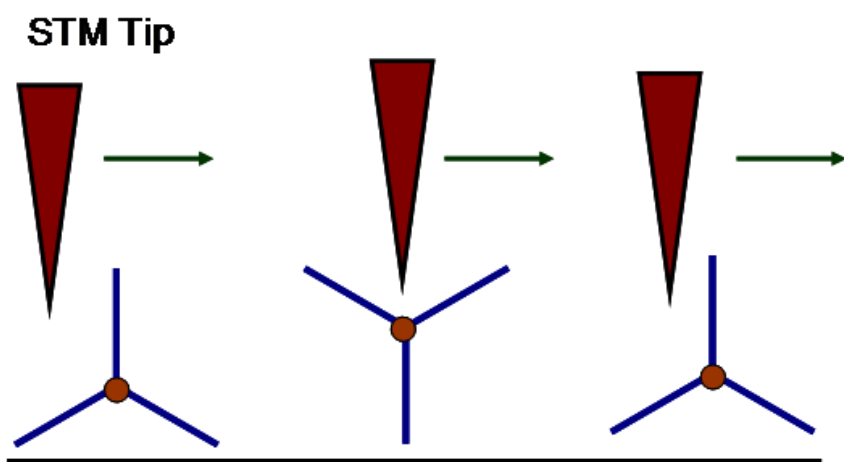
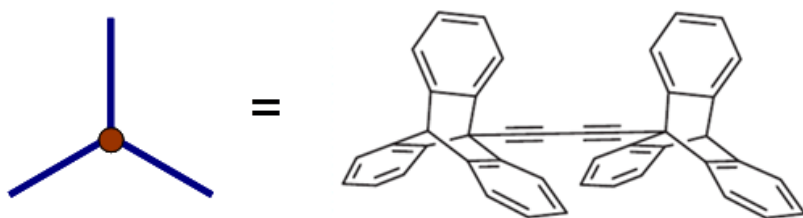
The STM imaging of C_{60} was first carried out at At King College, London. Similar to the atom positioning experiment by IBM, STM tip manipulation was also utilized to achieve C_{60} displacement. The tip trajectory suggested that a rolling motion took into account the displacement on the surface of C_{60} . In order to confirm the hypothesis, the researchers also employed *ab initio* density function (DFT) calculations with rolling model boundary condition ([\[link\]](#)). The calculation result has supported their experimental result.



Proposed mechanism of C₆₀ translation showing the alteration of C₆₀...surface interactions during rolling. a) 2-point interaction. The left point interaction was dissociated during the interaction. b) 1-point interaction. C₆₀ can pivot on surface. c) 2-point interaction. A new interaction formed to complete part of the rolling motion. a) - c) The black spot on the C₆₀ is moved during the manipulation. The light blue Si balls represent the first layer of molecules the silicon surface, and the yellow balls are the second layer.

The results provided insights into the dynamical response of covalently bound molecules to manipulation. The sequential breaking and reforming of highly directional covalent bonds resulted in a dynamical molecular response in which bond breaking, rotation, and translation are intimately coupled in a rolling motion ([\[link\]](#)), but not performing sliding or hopping motion.

A triptycene wheeled dimeric molecule [\[link\]](#) was also synthesized for studying rolling motion under STM. This "tripod-like" triptycene wheel unlike a ball like C₆₀ molecule also demonstrated a rolling motion on the surface. The two triptycene units were connected via a dialkynyl axle, for both desired molecule orientation sitting on surface and directional preference of the rolling motion. STM controlling and imaging was demonstrated, including the mechanism [\[link\]](#).

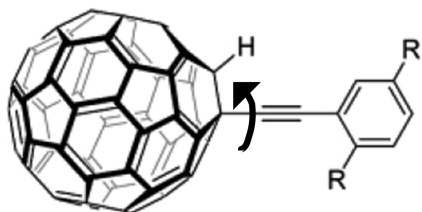


Scheme of the rolling mechanism (left to right). Step 1 is the tip approach towards the molecule, step 2 is a 120 degree rotation of a wheel around its molecular axle and in step 3 the tip reaches the other side of the molecule. It shows that, in principle, only one rotation of a wheel can be induced (the direction of movement is marked by arrows).

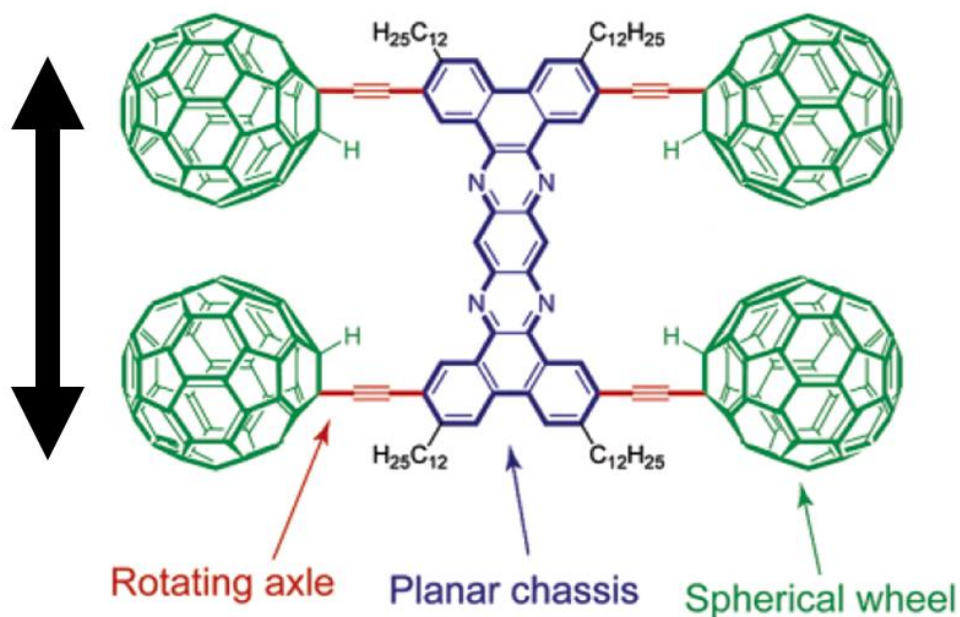
Single molecule nanocar under STM imaging

Another use of STM imaging at single molecule imaging is the single molecule nanocar by the Tour group at Rice University. The concept of a nanocar initially employed the free rotation of a C-C single bond between a spherical C_{60} molecule and an alkyne, [\[link\]](#). Based on this concept, an “axle” can be designed into which are mounted C_{60} “wheels” connected with a “chassis” to construct the “nanocar”. Nanocars with this design are

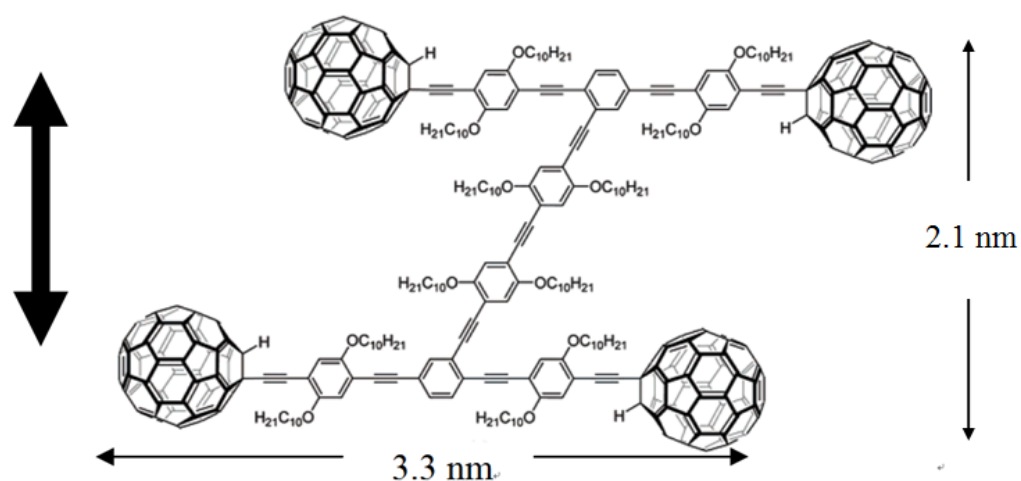
expected to have a directional movement perpendicular to the axle. Unfortunately, the first generation nanocar (named “nanotruck” [\[link\]](#)) encountered some difficulties in STM imaging due to its chemical instability and insolubility. Therefore, a new of design of nanocar based on OPE has been synthesized [\[link\]](#).



Structure of C_{60} wheels connecting to an alkyne. The only possible rolling direction is perpendicular to the C-C single bond between C_{60} and the alkyne. The arrow indicates the rotational motion of C_{60} .

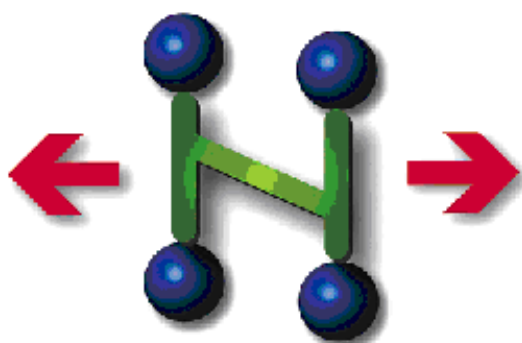


Structure of the nanotruck. No rolling motion was observed under STM imaging due to its instability, insolubility and inseparable unreacted C₆₀. The double head arrow indicates the expected direction of nanocar movement. Y. Shirai, A. J. Osgood, Y. Zhao, Y. Yao, L. Saudan, H. Yang, Y.-H. Chiu, L. B. Alemany, T. Sasaki, J.-F. Morin, J. M. Guerrero, K. F. Kelly, and J. M. Tour, *J. Am. Chem. Soc.*, 2006, **128**, 4854. Copyright American Chemical Society (2006).

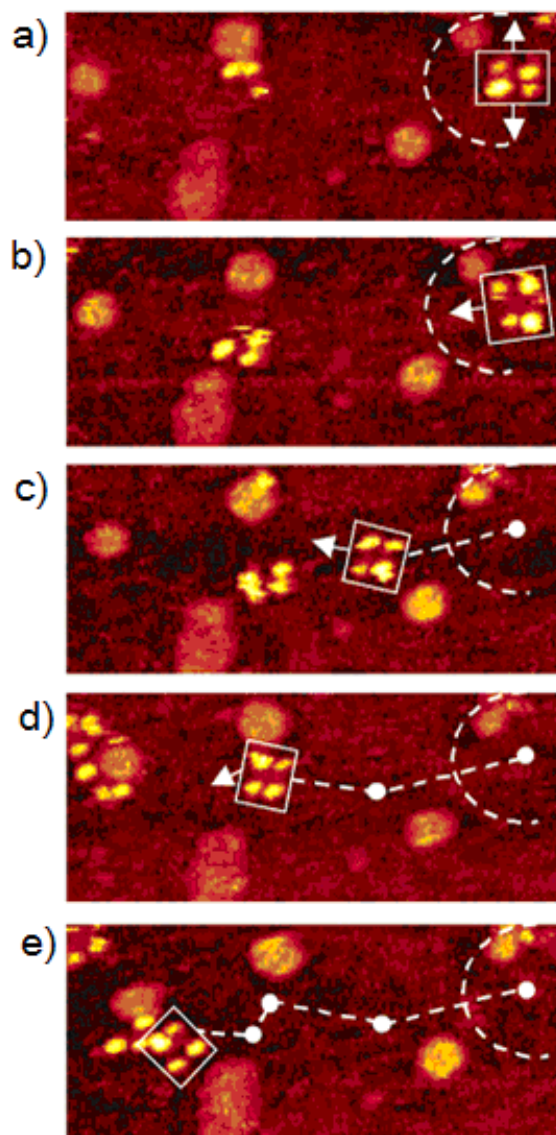


Nanocar based on OPE structure. The size of the nanocar is 3.3 nm X 2.1 nm (W x L). Alkoxy chains were attached to improve solubility and stability. OPE moiety is also separable from C₆₀. The bold double head arrow indicates the expected direction of nanocar movement. The dimension of nanocar was 3.3 nm X 2.1 nm which enable direct observation of the orientation under STM imaging. Y. Shirai, A. J. Osgood, Y. Zhao, K. F. Kelly, and J. M. Tour, *Nano Lett.*, 2005, 5, 2330. Copyright American Chemical Society (2005).

The newly designed nanocar was studied with STM. When the nanocar was heated to ~200 °C, noticeable displacements of the nanocar were observed under selected images from a 10 min STM experiment [\[link\]](#). The phenomenon that the nanocar moved only at high temperature was attributed their stability to a relatively strong adhesion force between the fullerene wheels and the underlying gold. The series of images showed both pivotal and translational motions on the surfaces.



Translational Motion



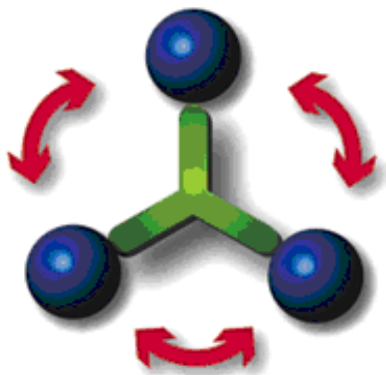
Pivotal and translational
movement of OPE based

nanocar. Acquisition time of one image is approximately 1 min with (a – e) images were selected from a series spanning 10 min. The configuration of the nanocar on surface can be determined by the distances of four wheels. a) – b) indicated the nanocar had made a 80° pivotal motion. b) – e) indicated translation interrupted by small-angle pivot perturbations. Y. Shirai, A. J. Osgood, Y. Zhao, K. F. Kelly, and J. M. Tour, *Nano Lett.*, 2005, 5, 2330. Copyright American Chemical Society (2005).

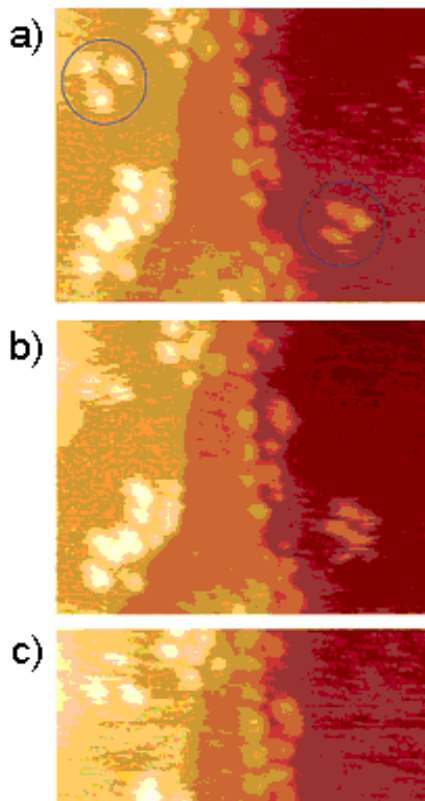
Although literature studies suggested that the C₆₀ molecule rolls on the surface, in the nanocar movement studies it is still not possible to conclusively conclude that the nanocar moves on surface exclusively via a rolling mechanism. Hopping, sliding and other moving modes could also be responsible for the movement of the nanocar since the experiment was carried out at high temperature conditions, making the C₆₀ molecules more energetic to overcome interactions between surfaces.

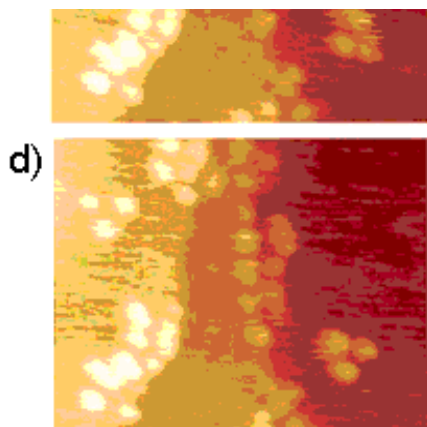
To tackle the question of the mode of translation, a trimeric “nano-tricycle” has been synthesized. If the movement of fullerene-wheeled nanocar was based on a hopping or sliding mechanism, the trimer should give observable translational motions like the four-wheeled nanocar, however, if rolling is the operable motion then the nano-tricycle should rotate on an axis, but not translate across the surface. The result of the imaging experiment of the trimer at ~200 °C ([\[link\]](#)), yielded very small and insignificant translational displacements in comparison to 4-wheel nanocar ([\[link\]](#)). The trimeric 3-

wheel nanocar showed some pivoting motions in the images. This motion type can be attributed to the directional preferences of the wheels mounted on the trimer causing the car to rotate. All the experimental results suggested that a C_{60} -based nanocar moves via a rolling motion rather than hopping and sliding. In addition, the fact that the thermally driven nanocar only moves in high temperature also suggests that four C_{60} have very strong interactions to the surface.



Pivoting Motion





Pivot motion of the trimer. a) - d) Pivot motions of circled trimers were shown in the series of images. No significant translation were observed in comparison to the nanocar. Y. Shirai, A. J. Osgood, Y. Zhao, K. F. Kelly, and J. M. Tour, *Nano Lett.*, 2005, 5, 2330. Copyright American Chemical Society (2005).

Bibliography

- D. M. Eigler and E. K. Schweizer, *Nature*, 1990, **344**, 524.
- L. Grill, K. -H. Rieder, F. Moresco, G. Rapenne, S. Stojkovic, X. Bouju, and C. Joachim, *Nat. Nanotechnol.*, 2007, **2**, 95.
- Y. Shirai, A. J. Osgood, Y. Zhao, K. F. Kelly, and J. M. Tour, *Nano Lett.*, 2005, **5**, 2330.

- Y. Shirai, A. J. Osgood, Y. Zhao, Y. Yao, L. Saudan, H. Yang, Y.-H. Chiu, L. B. Alemany, T. Sasaki, J.-F. Morin, J. M. Guerrero, K. F. Kelly, and J. M. Tour, *J. Am. Chem. Soc.*, 2006, **128**, 4854.

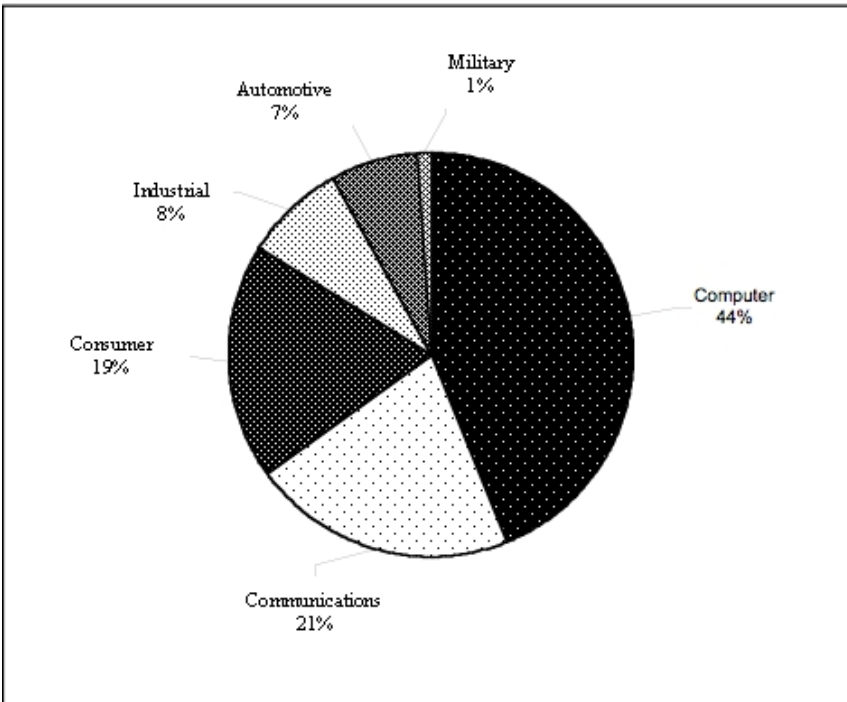
The Environmental Impact of the Manufacturing of Semiconductors
This module gives a brief general overview of semi-conductor manufacturing and some of the components and processes used to produce them that can potentially cause harm to humans or the environment.

Note: "This module was developed as part of a Rice University Class called "[Nanotechnology: Content and Context](#)" initially funded by the National Science Foundation under Grant No. EEC-0407237. It was conceived, researched, written and edited by students in the Fall 2005 version of the class, and reviewed by participating professors."

What is a semiconductor?

The semiconductor industry is one of the fastest growing manufacturing sectors in not only the United States but also in the world. According to the American Electronics Association, the domestic sales of electronic components have skyrocketed, jumping from \$127 billion to \$306 billion over the course of the 1980's. In the first three quarters of the 2003 fiscal year alone, the export of technology goods from the United States increased by \$19 billion [1].

The word "semiconductor" technically refers to any member of a class of solid, crystalline materials that is characterized by an electrical conductivity better than that of insulators (e.g., plastic) but less than that of good conductors (e.g., copper) [2]. Semiconductors are particularly useful as a base material in the manufacturing of computer chips, and the term semiconductor has actually come to be synonymous with the computer chips, themselves. However, semiconductors are not only used in computers. Computers only make up 44% of entire industry consumption (see [\[link\]](#)). Semiconductors are also used for military, automotive, industrial, communications, and other consumer purposes.

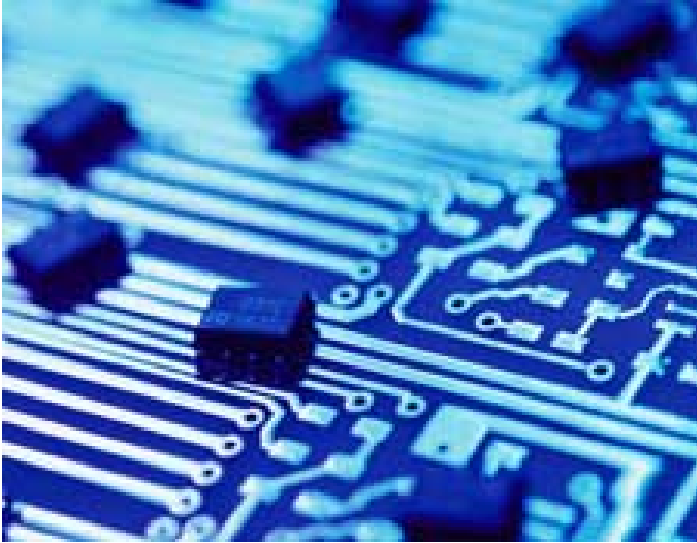


Relative consumption of semiconductors by industry [3].

Semiconductors seem to be anywhere and everywhere throughout our everyday lives, yet it is surprising how little most people know about how they actually work or about the potentially devastating effects their manufacturing can have on the environment and human health.

Why is nanotechnology important to the semiconductor industry?

Much of the study of nanotechnology has been centered on the manufacturing of semiconductors. Though there are a number of highly anticipated applications for nanotechnology in other fields, notably in medicine and in biotechnology, the most tangible results thus far can be argued to have been achieved in the semiconductor industry.



An example of a semiconductor
(photo from PEAK).

For example, Intel recently unveiled its first products based on a generation of 90-nanometer process technology, and its researches and engineers have built and tested prototype transistors all the way down to the 22-nanometer range. Currently, Intel scientists and engineers are working on identifying new materials such as carbon nanotubes and nanowires to replace current transistors, and in particular they hope to develop a “tri-gate” transistor approach that would enable chip designers to build transistors below the 22-nanometer range [4].

With the advent of nanotechnology, these transistors are becoming even faster and more powerful, and in accordance with the law of accelerating returns, the industry has been producing smaller transistors at lower costs with each and every passing year. As these semiconductors become smaller and smaller, they are quickly and surely pushing towards the limits of the nano-realm.

These innovations, however, do not come without their fair share of challenges. Physical issues that are not problematic at the micron scale arise at the nano-scale due to the emergence of quantum effects, and in much the

same way that optical microscopy cannot be utilized at the nano-scale, the semiconductor industry is fast approaching a similar diffraction limit. Optical lithography, for instance, a process that uses the properties of light to etch transistors onto wafers of silicon, will soon reach its limit.

At its most basic level, nanotechnology involves pushing individual atoms together one by one. Since approximately 1.7 billion transistors are required for a single chip, this is obviously not a realistic method for mass production. Unless an alternative method for production or a solution to this problem is found, the development and manufacturing of transistors are expected to hit a proverbial brick wall by the year 2015. This is the reason that research in nanotechnology is so important for the world and future of semiconductors.

How are semiconductors manufactured?

Today's semiconductors are usually composed of silicon, and they are manufactured in a procedure that combines the familiar with the bizarre; some steps that are involved in the process are as everyday as developing a roll of photographic film while others seem as if they would be better suited to take place on a spaceship.

These semiconductors appear to the naked eye as being small and flat, but they are actually three-dimensional "sandwiches" that are ten to twenty layers thick. It can take more than two dozen steps and up to two full months to produce a single one of these silicon sandwiches. Some of the basic and more essential steps involved in the manufacturing process of silicon chips are briefly detailed below.

First, silicon crystals are melted in a vat and purified to 99.9999% purity. The molten silicon is drawn into long, heavy, cylindrical ingots, which are then cut into thin slices called wafers about the thickness of a business card.

One side of each wafer must be polished absolutely smooth. This process is called chemical-mechanical polishing, and it involves bathing the wafers in special abrasive chemicals. After chemical-mechanical polishing,

imperfections cannot be detected on the wafers even with the aid of a laboratory microscope.

After a wafer is polished, layers of material must be stacked on top of the silicon wafer base. Insulating layers are laid down in alternation with conducting layers in a process called deposition. This is often achieved by spraying the chemicals directly onto the surface of the wafer through chemical vapor deposition. Following deposition, the wafer is coated with another layer of chemicals called a photoresist that is sensitive to light.

Next, a machine called a stepper ([link](#)) is calibrated to project an extremely fine and focused image through a special type of reticle film in a manner similar to that of a simple slide projector. The light that is transmitted through the reticle is projected onto the photoresist layer, which reacts to the light and begins to harden. All of the parts of the wafer exposed to this light harden into a tough crust while the parts in shadow remain soft. This particular step is known by the name of photoelectrochemical etching because it achieves an etching effect, resulting in a chip.



An artist's illustration of a stepper (image

from Solid State Electronics).

Hundreds of copies of the chip are etched onto the wafer until the entire surface has been exposed. Once this process is complete, the entire wafer is submerged into an etching bath, which washes away any parts of the photoresist that remain unexposed along with the insulating chemicals underneath. The hardened areas of the photoresist, however, remain and protect the layers of material underneath them. This process of depositing chemicals, coating with a photoresist, exposure to light over a film mask, and etching and washing away is repeated more than a dozen times. The result is an elaborate, three-dimensional construction of interlocking silicon wires.

This product is then coated with another insulating layer and is plated with a thin layer of metal, usually either aluminum or copper. Yet another photoresist is laid down on top of this metal plating, and after the wafer is exposed in a stepper, the process repeats with another layer of metal. After this step has been repeated several more times, a final wash step is performed, and a finished semiconductor product rolls off the assembly line, at last.

What is a clean room?

A typical semiconductor fabrication facility, or “fab” in industry jargon, looks like a normal two- or three-story office building from the outside, and most of the interior space is devoted to one or more “clean rooms,” in which the semiconductors are actually made. A clean room is designed with a fanatical attention to detail aimed towards keeping the room immaculate and dust-free ([link](#)).



An industry clean room at AP Tech (photo from Napa Gateway).

Most if not all surfaces inside these clean rooms are composed of stainless steel, and these surfaces are sloped whenever possible or perforated by grating to avoid giving dust a place to settle. The air is filtered through both the ceiling and the floor to remove particles that are down to 1/100 the width of a human hair. Lighting is characteristically bright and slightly yellowish to prevent mildew from forming behind equipment or in recessed corners, and even the workers in a clean room must be absolutely spotless.

Workers in these rooms must be covered from head to toe in “bunny suits” that completely seal the body in a bulky suit, helmet, battery pack, gloves, and boots. Once sealed in these suits, the workers often look more like space explorers in a science fiction movie than computer chip employees, but in order to even enter the stainless steel locker room to suit up to begin with, they must first pass through a series of air lock doors, stand under a number of “air showers” that actually blow dust off of clothing, and walk across a sticky floor matting that removes grime from the bottom of shoes.

Semiconductor-manufacturing companies often portray their fabrication facilities as being clean, environmentally friendly, and conspicuously free of the black, billowing smokestacks that have come to be associated with

the plants and factories of other major industries. These facilities produce no visible pollution and certainly do not appear to pose any health or environmental risks.

In truth, the term “clean room,” itself is more than just a bit of an understatement. Industry executives often boast that their clean rooms are from 1,000 times to 10,000 times cleaner and more sanitary than any hospital operating room.

What are the health risks involved in the semiconductor industry?

The use of sterile techniques and the fastidious attention devoted to cleanliness in the semiconductor industry may perpetuate the illusion that the manufacturing of semiconductors is a safe and sterile process. However, as a rapidly growing body of evidence continues to suggest, hardly anything could be further from the truth ([link](#)). The question of worker safety and chemical contamination at chip-making plants has received an increasing amount of attention over the course of the past decade.



Chemicals used in the
manufacturing of semiconductors
are known to have toxic effects
(image from FARSHA).

The devices being built at semiconductor fabrication facilities are super-sensitive to environmental contaminants. Because each chip takes dozens of trained personnel several weeks to complete, an enormous amount of time and effort is expended to produce a single wafer. The industry may pride itself on its perfectly immaculate laboratories and its bunny-suited workers, but it should be noted that the bunny suits are not designed to protect their wearers from hazardous materials but rather to protect the actual semiconductor products from coming into contact with dirt, hair, flakes of skin, and other contaminants that can be shed from human bodies. They protect the silicon wafers from the people, not the people from the chemicals.

Lee Neal, the head of safety, health, and environmental affairs for the Semiconductor Industry Association, has been quoted as saying, “This is an environment that is cleaner than an operating room at a hospital.” However, this boast is currently being challenged by industry workers, government scientists, and occupational-health experts across the country and worldwide.

Industrial hygiene has always been an issue in the semiconductor industry. Many of the chemicals involved in the manufacturing process of semiconductors are known human carcinogens or pose some other serious health risk if not contained properly. [\[link\]](#) lists ten of the hazardous chemicals most commonly used in manufacturing semiconductors along with their known effects on human health.

Chemical name	Role in manufacturing process	Health problems linked to exposure
Acetone	Chemical-mechanical polishing of silicon wafers	Nose, throat, lung, and eye irritation, damage to the skin, confusion, unconsciousness, possible coma
Arsenic	Increases conductivity of semiconductor material	Nausea, delirium, vomiting, dyspepsia, diarrhea, decrease in erythrocyte and leukocyte production, abnormal heart rhythm, blood vessel damage, extensive tissue damage to nerves, stomach, intestine, and skin, known human carcinogen for lung cancer
Arsine	Chemical vapor deposition	Headache, malaise, weakness, vertigo, dyspnea, nausea, abdominal and back pain, jaundice, peripheral neuropathy, anemia
Benzene	Photoelectrochemical	Damage to bone

	etching	marrow, anemia, excessive bleeding, immune system effects, increased chance of infection, reproductive effects, known human carcinogen for leukemia
Cadmium	Creates “holes” in silicon lattice to create effect of positive charge	Damage to lungs, renal dysfunction, immediate hepatic injury, bone defects, hypertension, reproductive toxicity, teratogenicity, known human carcinogen for lung and prostate cancer
Hydrochloric acid	Photoelectrochemical etching	Highly corrosive, severe eye and skin burns, conjunctivitis, dermatitis, respiratory irritation
Lead	Electroplated soldering	Damage to renal, reproductive, and immune systems, spontaneous abortion, premature birth, low birth

		weight, learning deficits in children, anemia, memory effects, dementia, decreased reaction time, decreased mental ability
Methyl chloroform	Washing	Headache, central nervous system depression, poor equilibrium, eye, nose, throat, and skin irritation, cardiac arrhythmia
Toluene	Chemical vapor deposition	Weakness, confusion, memory loss, nausea, permanent damage to brain, speech, vision, and hearing problems, loss of muscle control, poor balance, neurological problems and retardation of growth in children, suspected human carcinogen for lung and liver cancer
Trichloroethylene	Washing	Irritation of skin, eyes, and respiratory tract, dizziness,

		drowsiness, speech and hearing impairment, kidney disease, blood disorders, stroke, diabetes, suspected human carcinogen for renal cancer
--	--	---

Chemicals of concern in the semiconductor industry [5].

Several semiconductor manufacturers including National Semiconductor and IBM have been cited in the past for holes in their safety procedures and have been ordered to tighten their handling of carcinogenic and toxic materials.

In 1996, 117 former employees of IBM and the families of 11 workers who had died of cancer filed suit against the chemical manufacturers Eastman Kodak Company, Union Carbide Corporation, J. T. Baker, and KTI Chemicals, claiming that they had suffered adverse health effects as a result of exposure to hazardous chemicals on the job in the semiconductor industry [5]. The lawsuit was filed in New York, which prevented the employees from suing IBM directly. A separate group of former IBM workers who had developed cancer filed suit against the company in California, alleging that they had been exposed to unhealthy doses of carcinogenic chemicals over the past three decades. Witnesses who testified in depositions in the New York state court in Westchester County described how monitors that were supposed to warn workers of toxic leaks often did not function because of corrosion from acids and water. They also alleged that supervisors sometimes shut down monitors to maintain production rates. When they lodged complaints with senior officials in the company, they claim to have been told not to “make waves” [6]. Meanwhile, 70 female workers in Scotland sued National Semiconductor Corporation, another U.S.-based company, claiming that they, too, were exposed to carcinogens on the job.

These lawsuits and the resulting publicity prompted a groundbreaking study by the Health and Safety Executive, which commissioned a committee to investigate these allegations [7]. The committee found that there were indeed unusually high levels of breast and other kinds of cancer among workers at National Semiconductor's fabrication facility in Greenock, Scotland. The committee concluded that the company had failed to ensure that the local exhaust ventilation systems adequately controlled the potential exposure of employees to hydrofluoric acid and sulphuric acid fumes and to arsenic dust. These findings proved to be extremely embarrassing for the company and for the industry. According to an official statement released by Ira Leighton, acting regional administrator of the New England branch of the U.S. Environmental Protection Agency, "National Semiconductor is a big business that uses a large amount of harmful chemicals and other materials. Our hazardous waste regulations were created to properly monitor dangerous chemicals and prevent spills. In order for it to work, it is important businesses to comply with all of the regulations. When companies fail to do this they are potentially putting people – their employees and neighbors – at risk [8]. "

Moreover, a study of fifteen semiconductor manufacturers published in the December 1995 issue of the American Journal of Independent Medicine showed that women working in the so-called clean rooms of the semiconductor fabs suffered from a 14% miscarriage rate.



Protesters at a rally staged against
IBM (photo from San Francisco
Independent Media Center).

The main problem in prosecution is that the industry does not have a single overarching and definitive process for manufacturing, and it is difficult to pinpoint one particular compound as causing a certain health problem because some plants use as many as 300 chemicals. Also, many of the manufacturing processes take place in closed systems, so exposure to harmful substances is often difficult to detect unless monitored on a daily basis.

Executives and spokespeople for the semiconductor industry maintain that any chip workers' cancers and other medical problems are more likely due to factors unrelated to the job, such as family history, drinking, smoking, or eating habits. They also say that over the years, as awareness of chemical hazards has grown, they have made efforts to phase out toxic chemicals and to lower exposure to others. They insist that they use state-of-the-art process equipment and chemical transfer systems that limit or prevent physical exposure to chemicals and point out that the substances used in the semiconductor industry are used in other industries without a major health or safety problem.

What environmental risks are involved?

In theory, attention to cleanliness is in the manufacturer's best interest not only from a health perspective but also from an economic. Many chemicals used in the production process are not expensive in and of themselves; however, the cost of maintaining these materials in an ultra-clean state can be quite high. This encourages the close monitoring of usage, the minimization of consumption, and the development of recycling and reprocessing techniques. Also, the rising costs of chemical disposal are prompting companies to conduct research into alternatives that use more environmentally friendly methods and materials. Individual companies and

worldwide trade associations were active in reducing the use and emission of greenhouse gases during the 1990's, and the industry as a whole has substantially reduced emissions over the last twenty years.

Nonetheless, there has been a history of environmental problems linked to the industry in Silicon Valley and other technology centers. To begin with, a tremendous amount of raw materials is invested in the manufacturing of semiconductors every year.

Moreover, a typical facility producing semiconductors on six-inch wafers reportedly uses not only 240,000 kilowatt hours of electricity but also over 2 million gallons of water every day [9]. Newer facilities that produce eight-inch and twelve-inch wafers consume even more, with some estimates going as high as five million gallons of water daily. While recycling and reusing of water does occur, extensive chemical treatment is required for remediation, and in dry or desert areas such as Albuquerque, New Mexico, home to plants for Motorola, Philips Semiconductor, Allied Signal and Signetics, Intel, and other high-tech firms, the high consumption of water necessary for the manufacturing of semiconductors can pose an especially significant drain on an already scarce natural resource [10]. The existence of economic mainstays including the mining industry and the established presences of Sandia National Laboratories and the Los Alamos National Laboratory make New Mexico an attractive location for high-tech tenants. However, the opening of fabrication facilities in the state leaves its farmers and ranchers in constant competition with the corporations for rights to water consumption. On average, the manufacturing of just 1/8-inch of a silicon wafer requires about 3,787 gallons of wastewater, not to mention 27 pounds of chemicals and 29 cubic feet of hazardous gases [11].



A community near Sutter Creek, California that has been designated as an EPA Superfund site as a result of arsenic contamination (photo from Alexander, Hawes, & Audet).

Contamination has also been an issue in areas surrounding fabrication plants. Drinking water was found to be contaminated with trichloroethane and Freon, toxins commonly used in the semiconductor industry, in San Jose, California in 1981 [12]. These toxins were later suspected to be the cause of birth defects of many children in the area. The culprits were Fairchild Semiconductor and IBM. The companies' underground storage tanks were found to have leaked tens of thousands of gallons of the toxic solvents into the ground. There are a number of semiconductor-related EPA cleanup sites in Silicon Valley, and there have been concerns raised about the cumulative air and groundwater pollution in Silicon Valley, as well.

Another area of concern is the eventual fate of discarded electronic systems such as computers, pagers, mobile phones, and televisions that contain semiconductor devices. Personal computers in particular are especially problematic because they become obsolete fairly rapidly and lose almost all of their market value within five or ten years after their date of manufacture. Tens of millions of PC's are sold in the United States each year, and they pose an environmental risk not only through their sheer bulk

in city dumps and landfills but also because their semiconducting devices often contain significant amounts of heavy metals, including lead and other potentially hazardous substances.

Why don't we hear more about this on the news?

Across the United States, approximately 60% of the manufacturing facilities for semiconductor devices are located in six states. These states listed in descending order are California, Texas, Massachusetts, New York, Illinois, and Pennsylvania. The industry appears to be concentrated in these particular locations in part because they are near the primary users, transportation routes, and experts in the field, but people of all ages in all fifty states are impacted by semiconductor technology. Consumerism of semiconductor products is only expected to increase in coming years. Apple, for instance, expects to have sold 23.6 million iPods, devices that rely on semiconductor technology, by the year 2006.

If semiconductors are so ubiquitous in our day-to-day lives, why is there so little awareness about the serious environmental and health risks that are involved in their manufacturing process? Part of the problem is that little is known about the long-term health or environmental consequences of exposure to the chemicals that are used in the process. Because the semiconductor industry is still relatively new, not many studies have been conducted on this topic, and existing data is often inconclusive. This being said, some scientists predict that the cancer rate in the silicon chip industry will rise significantly in the future because cancer can take as long as 20-25 years to manifest itself in populations of exposed workers.

The EPA does have regulations in effect that are aimed toward the purpose of controlling the levels of contaminants released and minimizing human and environmental exposure to them. However, current regulations do not mandate that American companies report on offshore manufacturing. Therefore, even as media coverage and general awareness increase, companies can simply outsource more and more of their fabrication facilities to, for example, Southeast Asia. Some companies, in fact, have begun to do so, and there have even already been studies conducted on the

health issues of workers in the electronics and semiconductor industries of Singapore and Malaysia [13].

Thus, changes in how and where semiconductor firms manufacture chips currently outstrip the present ability of the United States government and media institutions to track and monitor their potential threats to humans and the environment. If this situation is to change for the better in the near future, it is clear that radical reforms will need to take place on a number of different levels. However, the who, what, when, where, and why, so to speak, of that reform remains to be addressed.

Discussion questions

- How many electronics products do you use on a day-to-day basis? How many of these products contain semiconductors?
- Who do you think is ultimately responsible for initiating reform? The government? The corporation? The consumer?
- Do you think that the health and environmental incidents related to semiconductor manufacturing will remain isolated incidents? Or do you think that these incidents will become epidemic in the future?
- Do you think that nanotechnology will help the problem or make the problem worse?

Endnotes

1. M. Kazmierczak and J. James. Industry Data & Publications: U.S. High-Tech Exports, 2000-2004. 16 Nov. 2004. American Electronics Association. 17 Oct. 2005<http://www.aeanet.org/Publications/idjl_ushightechexports1204.asp>.
2. J. Turley, The Essential Guide to Semiconductors. Upper Saddle River, New Jersey: Prentice Hall Professional Technical Reference, 2003.
3. J. Turley, The Essential Guide to Semiconductors. Upper Saddle River, New Jersey: Prentice Hall Professional Technical Reference, 2003. From [Prentice Hall](#)
4. IBM Research Nanotechnology Homepage. IBM. 16 Oct. 2005 <<http://domino.research.ibm.com/comm/research.nsf/pages/r.nanotech>.

html>.

5. R. Chepesiuk, "Where the Chips Fall: Environmental Health in the Semiconductor Industry." *Environmental Health Perspectives* 107 (1999): 452-457.
6. Richards, "Industry Challenge: Computer-Chip Plants Aren't as Safe And Clean As Billed, Some Say – Women at Scottish Factory Tell of Spills and Fumes, Face Host of Medical Ills – Firms Won't Help Do a Study." *Wall Street Journal* 5 Oct. 1998, eastern ed.: A1.
7. A. Heavens, Chip Plants Take Heat For Toxics. 14 Jan. 2003. *Wired News*. 13 Oct. 2005,
<<http://www.wired.com/news/technology/0,1282,57191,00.html>>
8. M. Merchant, *Maine Semiconductor Plant Fined For Hazardous Waste Violations*. Boston: U.S. Environmental Protection Agency, Press Office, 2001.
9. P. Dunn, *Cleanliness Outside, Some Issues Outside*. 2 Oct. 2000. The Foundation for American Communications. 13 Oct. 2005
<http://www.facsnet.org/tools/sci_tech/tech/community/environ2.php3>
10. J. Mazurek, *Making Microchips: Policy, Globalization, and Economic Restructuring in the Semiconductor Industry*. Cambridge, Massachusetts: MIT Press, 1999.
11. C. Hayhurst, "Toxic Technology: Electronics and the Silicon Valley." *E: the Environmental Magazine* May-Jun. 1997: 4.
12. B. Pimentel, "The Valley's Toxic History." *San Francisco Chronicle* 30 Jan. 2004, final ed.: B1.
13. V. Lin, *Health, Women's Work, and Industrialization: Semiconductor Workers in Singapore and Malaysia*. New York: Garland Publishing, Inc., 1991.